



UNIVERSIDAD AUTONOMA DE MADRID
FACULTAD DE PSICOLOGIA

Propiedades y aplicaciones del método de las
distancias mínimo-cuadráticas (LSDM) para la
validación y análisis de atributos cognitivos

Memoria para optar al título de Doctora en Psicología
presentada por

Sonia Janeth Romero Martínez

Dirigida por
Vicente Ponsoda Gil

Mayo de 2010

UNIVERSIDAD AUTONOMA DE MADRID
FACULTAD DE PSICOLOGIA



Propiedades y aplicaciones del método de las
distancias mínimo-cuadráticas (LSDM) para la
validación y análisis de atributos cognitivos

TESIS DOCTORAL

Sonia Janeth Romero Martínez

Mayo de 2010

Agradecimientos

La matemática es la ciencia del orden y la medida, de bellas cadenas de razonamientos, todos sencillos y fáciles.

René Descartes (1596-1650).

La realización de la presente tesis doctoral no hubiera sido posible sin la ayuda sincera y desinteresada de muchas personas que han sido un soporte y apoyo muy importante en todo el proceso. Quiero agradecer a los doctores Dimiter Dimitrov y Jimmy de la Torre por sus valiosos aportes y sugerencias para el desarrollo de la tesis. He tenido mucha suerte al poder contar con el apoyo y soporte directo de los autores originales de los modelos que he abordado a lo largo del trabajo.

Un agradecimiento muy sincero y especial a mi director, Dr. Vicente Ponsoda. Sin su guía y constante apoyo este trabajo no hubiera sido posible. Espero haber correspondido a la confianza depositada en mí y a la firme convicción que siempre ha tenido sobre la calidad de nuestro trabajo. De igual manera, mi más sincero agradecimiento a los profesores del departamento de Psicología Social y Metodología de la Universidad Autónoma de Madrid, especialmente a Carmen Ximénez, por su apoyo en las fases iniciales del proceso, a Paco Abad y Julio Olea por sus comentarios, sugerencias y aportes en las diversas ocasiones en que he presentado los estudios de la tesis en los seminarios del área de metodología.

Doy gracias también a mi familia porque a pesar de la distancia sé que han estado siempre pendientes de mi bienestar y del desarrollo de mis estudios. A mis padres Tito y Clemencia, por el apoyo moral y por estar siempre a mi lado, a mis hermanas por toda su ayuda; Laura, gracias por la rutina en C++ y a Natalia por digitalizar los datos, ya sé que es mucho trabajo pero deseo que les sea recompensado hoy y siempre con muchas bendiciones. Gracias a mis tios Gladys y Miguel, a mi sobrina Cata, a Yolanda, Hernando y a todos los demás familiares que han estado pendientes de mí, porque el ánimo y apoyo que me han brindado siempre me ha dado fuerzas para seguir adelante.

Al Ministerio Español de Educación y Ciencia por la financiación de estos años en la senda del conocimiento y de las dos estancias en George Mason University en las cuales aprendí mucho.

A los directores, profesores de matemáticas y alumnos de séptimo grado de secundaria

de los colegios San Francisco I la Casona I.E.D y de la Institución Educativa Integrada de Fontibón en Bogotá, Colombia, quienes desinteresadamente decidieron participar en el estudio.

A los amigos, que siempre han estado allí. A Elsy por ayudarme a decantar el problema con nuestras largas conversaciones sobre el papel de la cognición en los procesos de medida. A Clara y Andrés por hablarme de LaTeX y enseñarme sus ventajas. A Fabiola, Graciela y Esther por estar siempre pendientes del desarrollo de la tesis, a Pablo por los múltiples cafés compartidos conversando sobre metodología, a Rocío y Romelia por ayudarme a establecer el contacto con los colegios donde se realizó el estudio empírico y a Sofía, Julián, Tavo y Ana por su apoyo y cariño.

A mis compañeros del aula de becarios, y a los que hoy ya son doctores, por su apoyo y ánimo en cada etapa que se pasa a lo largo de estos años de estudio.

Deseo especialmente dar gracias a mi compañero de camino, Xavier Ordóñez, por las horas de angustias, sacrificios y también de alegrías cuando las cosas iban por buen camino, por las largas discusiones sobre el papel de los atributos en la medida, por la edición de la tesis, por ayudarme con el LaTeX y las referencias en BibTex y especialmente por programar las rutinas de R. Gracias por acompañarme a los congresos del área de metodología, de la NCME - AERA y por animarme a presentar los trabajos de la tesis en diversos escenarios. Por el cariño y compañía en los momentos más difíciles y por el apoyo constante en el largo camino que hemos emprendido hace ya 6 años. Finalmente quiero agradecer y dedicar esta tesis a mi hijo Angel Daniel, por su paciencia, alegría y por iluminar mi vida. En general quisiera agradecer a todas las personas que han vivido conmigo la realización de esta tesis doctoral, por su apoyo, cariño y colaboración.

Resumen

Se han desarrollado diversos modelos de componentes cognitivos para la medición de atributos subyacentes a la ejecución de los ítems de un test. En el presente trabajo se han estudiado las propiedades del método de las distancias mínimo-cuadráticas (en inglés, LSDM; Dimitrov 2007) para el análisis y validación de atributos cognitivos. En dicho método se utilizan los parámetros de los ítems, estimados con alguno de los modelos de la teoría de respuesta al ítem, y la matriz de especificación cognitiva (\mathbf{Q}) para estimar, mediante una solución mínimo cuadrática, la probabilidad de dominio de los atributos en niveles fijos de habilidad. Un segundo objetivo de la presente tesis ha sido ilustrar las aplicaciones del método con datos reales.

La tesis se compone de tres estudios de simulación y un estudio empírico en el que se incluyen diversas aplicaciones del método. En el primer estudio de simulación se ha analizado la distribución del estadístico MAD, (en inglés Mean Absolute Difference) para evaluar la calidad de la recuperación de las Curvas Características de los Items (CCI). En dicho estudio también se han determinado puntos de corte para los valores MAD en condiciones en las que se manipula el tamaño de la prueba, el número de atributos involucrados y la dificultad relativa de los mismos. En el segundo estudio de simulación se han descrito los efectos de la introducción de diversos tipos de error de especificación en una matriz \mathbf{Q} de 4 atributos y 15 ítems. El objetivo del tercer estudio de simulación ha sido analizar la sensibilidad y especificidad de los criterios MAD y LSD (en inglés, Least Squares Distance) para la detección de 14 tipos de error de especificación de la matriz \mathbf{Q} en cuatro condiciones de diseño en las que se manipula el tamaño del test, la cantidad de atributos y su dificultad relativa.

Los estudios de aplicación tienen varios objetivos: el primero ha sido construir un test de operaciones aritméticas básicas (adición y sustracción) entre números enteros, que incorpora principios cognitivos en su diseño. La prueba se ha basado en una estructura de seis atributos identificados tras un estudio sistemático de los procesos de solución de este tipo de tareas. El segundo objetivo ha sido aplicar el LSDM para acopiar evidencia sobre la validez de la estructura cognitiva de la prueba y proponer tres nuevas herramientas para complementar los criterios de validación planteados por Dimitrov (2007). La tercera

finalidad de este trabajo ha sido la comparación de los resultados del LSDM con los obtenidos en otros dos modelos componenciales. El último objetivo ha sido proponer y probar una metodología para usar el LSDM como herramienta para el diagnóstico cognitivo utilizando las probabilidades de dominio de los atributos estimadas con el método para clasificar a los examinados en estados de conocimiento.

Los resultados de la presente tesis han permitido conocer mejor las propiedades, potencialidades y aplicabilidad del LSDM. El método ha resultado ser una herramienta útil tanto para la validación de componentes cognitivos como para el diagnóstico de examinados.

Palabras Clave: Modelos de Diagnóstico Cognitivo, Teoría de Respuesta al Item, Validez, Test de Aritmética, Matriz **Q**.

Abstract

Several models of cognitive components has been developed with the aim of measure cognitive attributes underlying to correct answer of test items. The objective of the present dissertation is to study the properties of the Least Squares Distance Method (LSDM; Dimitrov, 2007) for the validation and analysis of cognitive attributes. The referred method uses the item parameters, estimated with any item response theory model, and the cognitive specification represented in the **Q**-matrix to estimate (by means of least squares solution) the probability of attribute mastery in fixed ability levels. A second aim for the present dissertation is to illustrate the LSDM applications with real data.

The thesis is composed of three simulation studies and one empirical study with several applications of the method. The aim of the first simulation study is to analyze the distribution of the Mean Absolute Difference (MAD) values, in order to evaluate the LSDM recovery of the Item Characteristic Curves (ICCs). A second goal of this study is to determine cutoff points for the MAD in different conditions, by manipulating the test size, number of attributes involved and relative difficulty of attributes. The second simulation study describes the effects of **Q**-matrix misspecifications in a cognitive structure of 4 attributes and 15 items. The objective of the third simulation study is to analyse the sensitivity and specificity of LSDM criteria (MAD and LSD) for the detection of 14 different types of misspecification introduced in the **Q**-matrix. Four design conditions was analysed by manipulating the test size, number of attributes involved and relative difficulty of the attributes.

The application studies have several objectives: firstly, an arithmetic test of basic operations between integers (addition and subtraction) was designed. The item design incorporates cognitive principles and is based in a structure of 6 attributes identified in a systematic analysis of the response process of this kind of tasks. The second aim is to applied the LSDM for accumulate evidence about the validity of the cognitive structure, also, three new validation criteria for the LSDM were proposed. The third objective has been to compare the results of the LSDM with other componential models. The last goal was to propose and probe a methodology for use the LSDM as method for cognitive diagnosis by means of the classification of examinees in knowledge states using the attribute

probabilities estimated for each ability level.

The results of the present dissertation has permitted to better know the properties, potential and applicability of the LSDM. The method has resulted to be a useful tool both for the cognitive component validation and examinees diagnosis.

Key words: Cognitive Diagnosis Models, Item Response Theory, Validity, Arithmetic Test, **Q**-Matrix.

Índice

1. Introducción	1
1.1. La evaluación diagnóstica cognitiva	1
1.2. Conjunción entre Psicometría y Psicología Cognitiva	2
1.3. Modelos componenciales	4
1.4. La estructura cognitiva y matriz Q	5
1.5. Fases del diagnóstico cognitivo	7
1.6. Proceso de validación de la estructura cognitiva	8
1.6.1. Método de la distancia minimocuadrática (LSDM)	11
1.7. Problema y objetivos	13
1.7.1. Estructura de la tesis	13
1.7.2. Estudios de simulación	13
1.7.3. Estudio de aplicación con datos reales	16
2. Marco Teórico	19
2.1. Modelos basados en la TRI	19
2.1.1. Modelos precursores	20
2.1.1.1. Modelo de Suppes	20
2.1.2. Modelos Unidimensionales	20
2.1.2.1. Modelo LLTM	20
2.1.2.2. Modelo ALM	22
2.1.2.3. Modelo MIRID	23
2.1.2.4. Modelo 2PL-CM	24
2.1.3. Modelos Multidimensionales	25

2.1.3.1.	Modelo MLTM	25
2.1.3.2.	Modelo GLTM	26
2.1.4.	Modelos para la medición de estrategias	27
2.1.4.1.	Modelo de Mislevy y Verhest (1990)	27
2.1.4.2.	Modelo MLTM-E	28
2.2.	Modelos de Diagnóstico Cognitivo	29
2.2.1.	Modelo Rule Space	31
2.2.2.	Modelo DINA	34
2.2.3.	Modelo NIDA	36
2.2.4.	Modelo Unificado	37
2.2.5.	Modelos RUM - Fusión	39
2.2.6.	Modelo GDM	40
2.2.7.	Redes Bayesianas	42
2.3.	Método LSDM	44
2.4.	Otros métodos de validación	51
2.4.1.	Procedimientos empleados en la Psicología Cognitiva	51
2.4.2.	Procedimientos basados en modelos componenciales	51
2.4.3.	Procedimientos de validación independientes del modelo componencial	54
2.4.3.1.	Quadratic Assignment (QA)	54
2.4.3.2.	Análisis Clúster	56
2.4.3.3.	Ecuaciones estructurales (SEM)	59
2.5.	Implicaciones del enfoque psicométrico cognitivo	61
2.5.1.	Validez y diseño de pruebas	61
2.5.2.	Establecimiento de vínculos entre Psicología Cognitiva y Psicometría	65
2.5.3.	Características de los atributos	67
2.6.	Modelos cognitivos sobre aritmética	68
2.6.1.	Concepto y evolución de la habilidad aritmética	68
2.6.2.	Modelos cognitivos de aritmética mental	68
2.6.3.	Aritmética con números enteros	71
2.6.3.1.	Modelo de cuatro estrategias	71
2.6.3.2.	Modelo basado en la teoría de los esquemas	71

2.6.3.3. Modelos basados en la teoría de enseñanza-aprendizaje . . .	72
2.7. Modelo cognitivo considerado en la presente tesis	75
3. Análisis de la recuperación de las CCI's con el LSDM	77
3.1. Introducción	77
3.2. Procedimiento	78
3.2.1. Diseño	78
3.2.2. Pasos de la simulación	79
3.2.2.1. Simulación de las matrices Q	79
3.2.2.2. Simulación de las respuestas de los sujetos	80
3.2.2.3. Ajuste del modelo TRI	83
3.2.2.4. Aplicación del LSDM en las réplicas	83
3.3. Resultados	83
3.3.1. Ajuste del modelo TRI	83
3.3.2. Distribución del estadístico MAD	86
3.3.3. Análisis de la recuperación de las CCI's	86
3.4. Conclusiones	89
4. Descripción de los efectos de errores de especificación en la estructura cognitiva	91
4.1. Introducción	91
4.2. Método	94
4.2.1. Procedimiento	94
4.2.1.1. Errores introducidos en la matriz Q	94
4.2.1.2. Errores de heterogeneidad y falta de monotonicidad de los atributos	102
4.2.2. Métodos de análisis estadístico	104
4.3. Resultados	105
4.3.1. Errores de especificación en la matriz Q	105
4.3.1.1. Errores en los vínculos ítem-atributo	105
4.3.1.2. Errores de adición y fusión de atributos	113
4.3.1.3. Errores de dependencia e incompatibilidad entre atributos	121

4.3.2.	Errores de heterogeneidad y falta de monotonicidad	130
4.3.2.1.	Análisis de los efectos de la heterogeneidad	130
4.3.2.2.	Análisis de los efectos de la falta de monotonicidad	132
4.4.	Conclusiones	133
5.	Sensibilidad del LSDM para la detección de errores en Q	137
5.1.	Introducción	137
5.2.	Procedimiento	137
5.2.1.	Diseño del estudio	137
5.2.2.	Diseño de la simulación	138
5.2.2.1.	Generación de las matrices Q	138
5.2.2.2.	Generación de las matrices: $P(A = 1 \theta)$	139
5.2.2.3.	Generación de matrices P_{ij}	140
5.2.2.4.	Verificación de la calidad técnica del LSDM	140
5.2.2.5.	Obtención de réplicas de las matrices P_{ij}	141
5.2.2.6.	Introducción de errores de especificación en las matrices Q	141
5.2.2.7.	Aplicación del método LSDM	144
5.2.3.	Métodos de análisis estadístico	145
5.2.3.1.	Curvas ROC	145
5.2.3.2.	Análisis ROC de los valores LSDM y MAD	149
5.3.	Resultados	150
5.3.1.	Comparación entre condiciones	156
5.4.	Conclusiones	158
6.	Aplicaciones del LSDM	161
6.1.	Introducción	161
6.2.	Procedimiento	162
6.2.1.	Especificación de la estructura cognitiva	162
6.2.2.	Diseño y aplicación del Test de Aritmética	164
6.2.3.	Análisis mediante la TRI	167
6.2.4.	Aplicación del método LSDM	167
6.2.5.	Análisis de los atributos según su dificultad	167

6.2.6. Clasificación en estados de conocimiento	168
6.3. Métodos para evaluar la recuperación de las CCI	168
6.3.1. Diferencia Media Absoluta	168
6.3.2. Diferencia Media Absoluta Ponderada	168
6.3.3. Raíz Media de la Diferencia al Cuadrado	169
6.3.4. Método de límites.	169
6.4. El LSDM como herramienta para el diagnóstico cognitivo	170
6.5. Comparación de probabilidades de dominio	171
6.6. Resultados	172
6.6.1. Análisis según la Teoría Clásica de los Test	172
6.6.2. Aplicación del modelo de TRI	174
6.6.2.1. Dimensionalidad y aplicación del modelo de Rasch	174
6.6.3. Aplicación del LSDM	176
6.6.3.1. CPAs antes de la reespecificación:	176
6.6.3.2. CPAs después de la reespecificación:	177
6.6.4. Validación cruzada de los resultados con los modelos LLTM y DINA	182
6.6.4.1. Modelo LLTM	182
6.6.4.2. Modelo DINA	182
6.6.5. Comparación de los estados de conocimiento obtenidos con LSDM y DINA con los patrones observados	183
6.7. Conclusiones	184
7. Discusión General	187
8. Anexos	195
8.1. Anexo 1	195
8.2. Anexo 2	202
8.3. Anexo 3	206

Índice de Tablas

2.1. Matriz Q , estados de conocimiento (EC) y patrones de respuesta ideales de cada EC.	32
3.1. Número de patrones en Q para cada condición	80
3.2. Parámetros, error estándar y ajuste del modelo de Rasch en una réplica de 15 ítems	84
3.3. Parámetros, error estándar y ajuste del modelo de Rasch en una réplica de 24 ítems	84
3.4. Parámetros, error estándar y ajuste del modelo de Rasch en una réplica de 32 ítems	85
3.5. Matriz Q y percentil 95 en la condición 4A15I	87
3.6. Estadísticos de la distribución de valores MAD	88
3.7. Puntos de corte para los valores de MAD	89
4.1. Matriz Q verdadera con 4 atributos y 15 ítems.	94
4.2. Matrices Q esparcidas con errores de especificación	97
4.3. Matrices Q densas con errores de especificación	97
4.4. Matrices Q esparcidas con atributos añadidos o fusionados.	99
4.5. Matrices Q densas con atributos añadidos o fusionados	99
4.6. Matrices Q esparcidas con errores de dependencia	101
4.7. Matrices Q densas con errores de dependencia	101
4.8. Matrices Q esparcidas con errores de incompatibilidad	101
4.9. Matrices Q densas con errores de incompatibilidad	102
4.10. Parámetros b del atributo en cada ítem para la condición de heterogeneidad	104

4.11. Estadísticos MAD para evaluar la recuperación de la CCIs.	106
4.12. Valores residuales LSD en cada nivel fijo de habilidad	108
4.13. Estadísticos MAD para evaluar la recuperación de las CCIs	113
4.14. Valores residuales LSD en cada nivel fijo de habilidad	114
4.15. Estadísticos MAD para evaluar la recuperación de las CCIs	122
4.16. Valores residuales LSD en cada nivel fijo de habilidad	123
4.17. Valores MAD para cada ítem y LSD para cada nivel de habilidad	131
5.1. Distribución de ítems en las matrices \mathbf{Q} según el número posible de patrones	139
5.2. Algunas matrices \mathbf{Q} con errores de especificación de la condición 4A15I . .	143
5.3. Tabla de confusión para el estadístico MAD	146
5.4. Matriz de fracciones	146
5.5. Estadísticos ROC del criterio MAD en la condición 4A15I	151
5.6. Estadísticos ROC del criterio LSD en la condición 4A15I	152
5.7. Estadísticos ROC del criterio MAD en la condición 4A32I	153
5.8. Estadísticos ROC del criterio LSD en la condición 4A32I	153
5.9. Estadísticos ROC del criterio MAD en la condición 8A15I	154
5.10. Estadísticos ROC del criterio LSD en la condición 8A15I	154
5.11. Estadísticos ROC del criterio MAD en la condición 8A32I	155
5.12. Estadísticos ROC del criterio LSD en la condición 8A32I	155
6.1. Estructura cognitiva del Test de Aritmética	163
6.2. Estructura cognitiva del TA (matriz \mathbf{Q})	166
6.3. Estadísticos de la TCT para los ítems del TA	173
6.4. Dimensionalidad y ajuste al modelo de Rasch para los 39 ítems del TA . . .	175
6.5. Estadísticos de recuperación de las CCIs	179
6.6. Comparación del orden de la dificultad relativa de los atributos	183
6.7. Estadísticos MAD y RMSD para los modelos DINA y LSDM	184
8.1. Matriz \mathbf{Q} y percentil 95 en la condición 6A15I	195
8.2. Matriz \mathbf{Q} y percentil 95 en la condición 8A15I	196
8.3. Matriz \mathbf{Q} y percentil 95 en la condición 4A24I	196

8.4. Matriz Q y percentil 95 en la condición 6A24I	197
8.5. Matriz Q y percentil 95 en la condición 8A24I	198
8.6. Matriz Q y percentil 95 en la condición 4A32I	199
8.7. Matriz Q y percentil 95 en la condición 6A32I	200
8.8. Matriz Q y percentil 95 en la condición 8A32I	201
8.9. Matriz Q en la condición 4A15I	202
8.10. Matriz Q en la condición 4A32I	203
8.11. Matriz Q en la condición 8A15I	204
8.12. Matriz Q en la condición 8A32I	205
8.13. Test de Aritmética	206

Índice de Gráficas

2.1. Ejemplo de modelo de estudiante para tareas de sustracción de fracciones. . .	43
2.2. Ejemplo estrategia recta numérica	73
2.3. Ejemplo estrategia círculos	73
3.1. CPAs de las condiciones con 4 atributos	81
3.2. CPAs de las condiciones con 6 atributos	82
3.3. CPAs de las condiciones con 8 atributos	82
3.4. Distribución empírica del estadístico MAD	86
4.1. CPAs verdaderas generadas con el modelo de Rasch.	95
4.2. CPAs para las condiciones de falta de monotonicidad	103
4.3. CPAs condición esparcida, fácil, incluir (SFI)	108
4.4. CPAs condición esparcida, difícil, incluir (SDI)	109
4.5. CPAs condición esparcida, fácil, excluir (SFE)	110
4.6. CPAs condición densa, fácil, excluir (DFE)	110
4.7. CPAs condición esparcida, difícil, excluir (SDE)	112
4.8. CPAs condición densa, difícil, excluir (DDE)	112
4.9. CPAs de las condiciones de adición (SA y DA)	115
4.10. CPAs de la condición esparcida, fácil, AND, fusionar (SFAF)	116
4.11. CPAs de la condición densa, fácil, AND, fusionar (DFAF)	116
4.12. CPAs de la condición esparcida, difícil, AND, fusionar (SDAF)	117
4.13. CPAs de la condición densa, difícil, AND, fusionar (DDAF)	118
4.14. CPAs de la condición esparcida, fácil, OR, fusionar (SFOF)	119
4.15. CPAs de la condición densa, fácil, OR, fusionar (DFOF)	119

4.16. CPAs de la condición esparcida, difícil, OR, fusionar (SDOF)	120
4.17. CPAs de la condición densa, difícil, OR, fusionar (DDOF)	121
4.18. CPAs de la condición esparcida, fácil, dependencia simple (SFR1)	124
4.19. CPAs de la condición densa, fácil, dependencia simple (DFR1)	124
4.20. CPAs de la condición esparcida, difícil, dependencia simple (SDR1)	125
4.21. CPAs de la condición densa, difícil, dependencia simple (DDR1)	126
4.22. CPAs de la condición esparcida, fácil, incompatibilidad (SFI)	127
4.23. CPAs de la condición densa, fácil, incompatibilidad (DFI)	127
4.24. CPAs de la condición esparcida, difícil, incompatibilidad (SDI)	128
4.25. CPAs de la condición densa, difícil, incompatibilidad (DDI)	129
4.26. CPAs de la condición esparcida, dependencia compleja (SR2)	130
4.27. CPAs para la condición de heterogeneidad	132
4.28. CPAs para la condición de falta de monotonicidad	133
5.1. CPAs verdaderas para las condiciones con 4 atributos	139
5.2. CPAs verdaderas para las condiciones con 8 atributos	140
5.3. FVP de los valores MAD en cada condición de simulación	157
5.4. PAUC de los valores MAD en cada condición de error	158
6.1. Asignación de probabilidades de dominio de los atributos	171
6.2. Curvas de probabilidad de los atributos sin reespecificar \mathbf{Q}	176
6.3. Curvas de probabilidad de los atributos tras la reespecificación de \mathbf{Q}	177
6.4. CCI original y recuperada con límites para el ítem 2	180
6.5. CCI original y recuperada con límites para el ítem 8	181
6.6. CCI original y recuperada con límites para el ítem 16	181

Capítulo 1

Introducción

1.1. La evaluación diagnóstica cognitiva

El objetivo de la evaluación enfocada al diagnóstico cognitivo es medir las estructuras específicas de conocimiento y habilidades de procesamiento que subyacen a la ejecución de los ítems de un test. Por medio de este tipo de medición se busca proporcionar información sobre las fortalezas y debilidades cognitivas de los examinados. El interés por el desarrollo de modelos psicométricos que incorporan atributos, reglas o procesos cognitivos es relativamente reciente pero ha avanzado de forma vertiginosa, hasta el punto de que actualmente existe una gran diversidad de modelos. Esto puede deberse, en parte, a la necesidad creciente de hacer mediciones más informativas acerca de los procesos mentales que llevan a cabo las personas para resolver los ítems o tareas.

Los fundamentos teóricos que sustentan la medición diagnóstica cognitiva se basan en la integración de la Psicología Cognitiva y la Psicometría y tienen su origen en los capítulos de Messick (1989) y Snow y Lohman (1989). La investigación en el área es bastante activa en este momento, conduciendo a numerosas producciones académicas entre las que se pueden contar diversos artículos y libros recopilatorios como el editado por Nichols, Chipman y Brennan (1995) y el editado por Leighton y Griel (2007). Algunas de estas aproximaciones se han centrado en la implementación del diagnóstico cognitivo en la medición educativa, por ejemplo, a través de tutores inteligentes (e.g. Nichols, Chipman y Brennan, 1995), otros se han enfocado en las implicaciones de esta perspectiva en la validez de constructo y en el diseño de test (e.g. Leighton y Griel, 2007; Embretson, 1998, Nichols, 1994) y muchos

artículos se han dedicado a la formulación de modelos y a su aplicación (e.g Fischer, 1973; Tatsuoaka, 1983; Embretson, 1984; Junker y Sijtsma, 2001).

1.2. Conjunción entre Psicometría y Psicología Cognitiva

La unión entre Psicometría y Psicología Cognitiva se ha fortalecido con el desarrollo de esta última, que ha proporcionado herramientas para comprender la naturaleza, adquisición y uso del conocimiento humano, por medio del estudio de la mente en términos de las representaciones mentales y procesos que subyacen al comportamiento observable. Parte de la Psicología Cognitiva ha profundizado en el estudio del procesamiento de información, estrategias mentales, operaciones y procesos empleados por las personas cuando se enfrentan a problemas, tareas o ítems en el contexto de la medición. En particular, la teoría del procesamiento de información (van Lehn, 1989) ha proveído a la Psicometría de una base teórica para descomponer los ítems o tareas de los test en componentes cognitivos utilizados para su solución.

En cuanto al aporte de la Psicometría, la mayoría de los modelos de diagnóstico cognitivo se han basado en la Teoría de Respuesta al Ítem (TRI; Lord, Novick y Birnbaum, 1968). Dentro de la historia de la Psicometría, la TRI ha proporcionado las bases teóricas y estadísticas para llevar a cabo procesos de medición cada vez más precisos por medio del estudio de las propiedades de los ítems. Sin embargo, los modelos tradicionales de TRI proporcionan información general sobre el nivel de rasgo o habilidad de las personas pero no ofrecen un diagnóstico sobre los componentes de dicha habilidad. La unión entre Psicología Cognitiva y Psicometría ha permitido superar esta limitación de la TRI con una profundización en el diseño y estudio de los contenidos del ítem, de las estrategias de respuesta y de los atributos cognitivos que el ítem demanda.

La perspectiva psicométrico-cognitiva ha tenido también otras repercusiones positivas en la Psicometría tradicional entre las que se pueden contar: a) un desarrollo más eficiente de los ítems mediante la posibilidad de generar ítems automáticamente, b) el incremento en la validez de las inferencias que se realizan sobre las puntuaciones de la prueba pues desde la fase de diseño del test se tiene en cuenta la complejidad cognitiva subyacente y c) una retroalimentación más detallada de las fortalezas y dificultades de los examinados, ya

que estos modelos permiten interpretar los pasos intermedios ejecutados por las personas para resolver las tareas y no solo sus respuestas finales.

La primera ventaja mencionada anteriormente tiene que ver con la Generación Automática de Ítems (GAI). La GAI es un proceso mediante el cual se crean ítems de forma automática con propiedades psicométricas predecibles. Existen dos grandes aproximaciones para generar ítems de esta forma: la primera ha sido propuesta por Bejar (1993, 2002) y supone la creación de ítems isomorfos a partir de la estructura de ítems existentes. El segundo enfoque es el de Embretson (1993, 1998, 1999b). Para esta autora, la GAI es el producto final de la representación de constructo de un ítem, en otras palabras, los ítems se generan con base a sus características cognitivas subyacentes basadas en un modelo cognitivo del constructo a medir. La utilización de la GAI tiene múltiples ventajas, como una menor inversión en tiempo y dinero en la construcción de los bancos de ítems y un incremento en la seguridad de las pruebas derivado de una menor exposición de los ítems.

La segunda repercusión de este nuevo enfoque sobre la Psicometría tiene que ver con el diseño de las pruebas: la elaboración de los ítems desde este punto de vista se basa en una perspectiva centrada en el constructo (Messick, 1989, 1995; Embretson, 1993, 1998, 1999a) bajo la cual se busca entender y modelar los componentes subyacentes de los ítems para crearlos, seleccionarlos y ensamblarlos en el test de acuerdo a las características y componentes del constructo que se pretende medir. Esta forma de diseñar las pruebas implica una comprensión amplia de los atributos que componen el constructo e involucra el estudio profundo de los ítems antes del desarrollo del test, esto hace que las inferencias basadas en las puntuaciones de la prueba sean probablemente más válidas que con cualquier otro tipo de diseño. Sin embargo, como afirman Borsboom y Mellenbergh (2007): en sí mismo el diseño no garantiza que el test sea válido, la validez debe probarse mediante hipótesis que necesitan evidenciarse separadamente.

La tercera implicación mencionada tiene que ver con el tipo de retroalimentación ofrecida por este tipo de modelos. Desde una perspectiva psicométrico-cognitiva, la habilidad latente se divide en componentes elementales y se evalúa al examinado con respecto a su nivel de competencia en cada uno de ellos, esto permite una retroalimentación mucho más detallada, no solo del grado de dominio de cada uno de los componentes, sino incluso de los tipos de errores que los examinados son propensos a cometer.

1.3. Modelos componenciales

Los modelos derivados de la perspectiva psicométrico-cognitiva se han denominado de diversas maneras: modelos componenciales (Van der Linden y Hambleton, 1997), modelos de clasificación múltiple (Maris, 1999), modelos de clase latente restringida (Haertel, 1989), modelos TRI estructurados (Rupp y Mislevy, 2007), modelos de diagnóstico cognitivo (Nichols, Chipman y Brennan, 1995) y modelos de diagnóstico psicométrico (Rupp, 2007).

En el presente trabajo se adoptará el término modelos componenciales pues es una expresión general que no implica consideraciones sobre los parámetros, objetivos o técnicas empleados en el modelo. La mayoría de los modelos componenciales se han basado en la TRI y en los modelos de Clase Latente (CL) y se pueden dividir en dos grupos de acuerdo a sus objetivos: los modelos derivados de la TRI en los que se busca descomponer los parámetros de los ítems en atributos subyacentes y los modelos de diagnóstico cognitivo cuyo propósito es clasificar a los examinados en *estados de conocimiento* (Tatsuoka, 1995) de acuerdo al dominio que presentan los sujetos en cada uno de los atributos latentes.

Los modelos basados en la TRI pueden ser, a su vez, divididos en cuatro grupos:

1. Modelos *unidimensionales* que descomponen los parámetros de los ítems en componentes cognitivos; ejemplos de este grupo son el modelo lineal logístico de rasgo latente (LLTM; Fischer, 1973) y el modelo logístico de dos parámetros restringido (Embretson 1999a).
2. Modelos *multidimensionales* que descomponen la habilidad para resolver ítems dicotómicos o politómicos en atributos cognitivos; por ejemplo el modelo logístico atomatón (ALM; Spada, 1977), el modelo multicomponencial de rasgo latente (MLTM; Whitely, 1980), el modelo general de rasgo latente (GLMT; Embretson, 1984), el modelo con restricciones internas en los parámetros de dificultad (MIRID; Butter, De Boeck y Verhelst (1998) y el modelo de diagnóstico general (GDM; Von Davier y Yamamoto, 2004; Von Davier, 2008).
3. Modelos de *estrategias* de respuesta como el propuesto por Mislevy y Verhelst (1990) y por Embretson (1985).
4. Modelos *mixtos* que vinculan la TRI con los modelos de clase latente entre los que

se pueden encontrar el modelo MIRA (Rost, 1990) y el modelo híbrido (Yamamoto y Gitomer, 1993).

Los modelos de *diagnóstico cognitivo* buscan clasificar a los examinados en estados de conocimiento de acuerdo a su grado de dominio de los atributos cognitivos. Algunos ejemplos de esta aproximación son el *Rule Space Model* (RSM; Tatsuoka, 1983), los modelos *Deterministic Input, Noisy And Gate* (DINA; Junker y Sijtsma, 2001) y *Noisy Input, Deterministic And Gate* (NIDA; de la Torre y Douglas, 2004), los modelos Unificado y Fusión (DiBello, Stout y Roussos, 1995, Hartz, 2002) y el enfoque en el que se aplican las redes bayesianas (Mislevy, 1995).

1.4. La estructura cognitiva y matriz Q

Uno de los elementos claves de la medición de diagnóstico cognitivo es la especificación de la estructura cognitiva subyacente al test. Este proceso consiste en determinar el tipo, cantidad y relaciones entre los componentes o atributos cognitivos y establecer sus vínculos con ítems o tareas concretas. Los vínculos entre atributos e ítems se suelen representar en una matriz binaria llamada *matriz Q* (Tatsuoka, 1983) en la que j representa el ítem y k los atributos cognitivos. Los elementos q_{jk} especifican si el atributo A_k es necesario para resolver correctamente el ítem j ($q_{jk} = 1$) o no ($q_{jk} = 0$).

En el proceso de especificación de la estructura juega un papel fundamental el nivel de desarrollo de las teorías cognitivas; sin embargo, dichas teorías no pueden usarse directamente sino que deben adaptarse a los propósitos de la medición, en palabras de Leighton y Griel (2007): los psicómetras deben adaptar sus métodos, técnicas y herramientas para crear, modificar y probar las teorías cognitivas.

Uno de los aspectos más importantes en la adaptación de las teorías cognitivas a los procesos de medida es la decisión sobre el grado de especificidad con el que se pretende evaluar. En el marco de la Psicología Cognitiva se han construido detallados sistemas de producción que se han operacionalizado en tutores inteligentes como el ACT (Corbett, Andreson, O'Brien, 1995), el LISP (Anderson, 1990) o el QUEST (White y Fredericksen, 1987). También se han desarrollado complejas redes procedimentales, conceptuales o semánticas en diversas áreas del conocimiento (Naveh-Benjamin, Lin y McKeachie, 1995;

Johnson, Golsmith y Teague, 1995). Estos desarrollos se han basado en revisiones de la teoría mediante experimentos e investigaciones empíricas llevadas a cabo por los científicos cognitivos para comprender la forma en que las personas entienden, conceptualizan, razonan y resuelven problemas (Nichols, 1994, Lohman, 2000).

Según Yang y Embretson (2007), la medición diagnóstica tanto en Educación como en Psicología se centra principalmente en tres características cognitivas: a) perfiles de habilidades o listas de reglas y conocimientos, b) redes estructuradas de procedimientos y/o de conocimientos y c) procesos, componentes o capacidades cognitivas. Estas tres características no son exhaustivas y hay otras habilidades de pensamiento que también pueden ser incluidas en la medición diagnóstica; por ejemplo habilidades de alto orden como estrategias o procesos metacognitivos.

En las publicaciones del área, los autores han usado indistintamente diversos términos para referirse a los componentes cognitivos, algunos de ellos son: habilidades, estrategias, atributos, esquemas, rasgos, reglas de producción, competencias, operaciones y procesos cognitivos. Cada uno de dichos términos expresa diferentes niveles de cognición y una de las tareas del proceso de evaluación es establecer en cuál de estos niveles se pretende realizar la medición de acuerdo a los objetivos de la misma (Mislevy, 1995). En el presente trabajo se empleará la palabra general atributos para designar a los componentes, independientemente del nivel cognitivo en el que se pretende medir. Desde la perspectiva del modelamiento psicométrico la especificidad y el nivel de detalle de los elementos cognitivos en la medición suelen encontrarse limitados por la capacidad técnica de los modelos y de las herramientas informáticas disponibles. Algunos de los modelos propuestos han resultado inviables porque sus parámetros no se pueden estimar, como el modelo unificado de DiBello, Stout y Roussos (1995). La mayoría de modelos de diagnóstico cognitivo resultan afectados cuando se incluyen muchos atributos, tal es el caso del RSM (Tatsuoka, 1983), de los modelos DINA y NIDA (Junker y Sijtsma, 2001) y del modelo Fusión (Hartz, 2002). En cuanto a las herramientas informáticas, en los tutores inteligentes más complejos se usan técnicas bayesianas para realizar inferencias sobre la habilidad de los examinados, sin embargo, la complejidad computacional requerida por algunos sistemas con cientos de reglas puede ser excesiva generando problemas de estimación (Mislevy, 1995).

1.5. Fases del diagnóstico cognitivo

El desarrollo de un proceso evaluativo de orientación psicométrico-cognitiva debe seguir las siguientes fases:

1. Construcción de la teoría sustantiva. Este paso requiere el desarrollo o adaptación de un modelo teórico que caracterice las estructuras de conocimiento hipotéticas y los procesos requeridos en la medición. También se deben seleccionar las variables específicas mediante las que se puede acceder a dichos procesos o estructuras. En otras palabras, en esta fase se determina la cantidad, tipo y relaciones de los atributos.
2. Diseño de las tareas y especificación de vínculos entre atributos e ítems. En este paso se crean o seleccionan los ítems y tareas que corresponden con los procesos y estructuras de conocimiento identificados en la fase anterior y se realiza la especificación de estructura cognitiva del test.
3. Administración de la prueba. Este paso involucra importantes detalles del ambiente y contexto en el que se realizará la medición. Algunos de ellos son la selección del formato de los ítems, la forma de presentación, su organización en el test, entre otros. Es importante que el investigador tenga en cuenta la forma en que estas variables pueden afectar la ejecución de los sujetos.
4. Selección y aplicación de un modelo componencial. En esta fase el investigador elige y aplica, de acuerdo a sus necesidades y a las características del modelo teórico, alguno de los modelos matemáticos disponibles para estimar la probabilidad de acierto de cada ítem de acuerdo a las propiedades estructurales de los mismos y a los componentes de habilidad de los examinados. Para la selección apropiada de un modelo hay que tener en cuenta variables como el tipo de relación entre los atributos, nivel de especificidad, calidad de la estructura teórica disponible, formato de los ítems y objetivos de la medición, entre otras. Por ejemplo, dependiendo del tipo de relación que puede existir entre los atributos se elegirá un modelo no compensatorio, en el que el dominio de un atributo no compensa la carencia en otros atributos, o por el contrario, un modelo compensatorio en el que el dominio de un atributo subsana la insuficiencia en otros componentes.

5. Puntuación y retroalimentación de las respuestas. Este paso incluye asignar puntuaciones a los examinados que sean informativas sobre los componentes que se están evaluando a través del test. Estas puntuaciones permiten retroalimentar tanto a profesores como a alumnos en el diseño de estrategias de mejoramiento de los procesos de enseñanza-aprendizaje enfocadas a las necesidades individuales de cada estudiante. En este punto es necesario tomar decisiones sobre la forma de retroalimentación que se va a ofrecer; por ejemplo, esta puede ser inmediata y enfocada al mejoramiento del desempeño en la tarea como en el caso de los tutores inteligentes.
6. Revisión de la estructura cognitiva: El último paso consiste en reexaminar el diseño estructural de la medición para observar si los resultados de la misma apoyan el modelo teórico en el que se ha basado, en caso de no ser así, se reconsidera el diseño y se inicia nuevamente todo el proceso, ya que si los atributos propuestos no corresponden a aquellos que realmente utilizan los examinados, los resultados de la medición pueden ser erróneos y no proporcionarán información diagnóstica precisa, significativa y útil. Es en esta fase donde el método de las distancias mínimo-cuadráticas (LSDM, Dimitrov, 2007) tiene una especial relevancia.

1.6. Proceso de validación de la estructura cognitiva

El concepto de validez ha evolucionado mucho desde la elaboración de Cronbach y Meehl en 1955. Los resultados más importantes de dicha evolución han sido la integración de varios aspectos de la validez en un marco general (Messick, 1989, 1995) y una conceptualización diferente de la validez que se refleja en los últimos estándares para la medición psicológica y educativa (American Education Research Association, American Psychological Association and National Council on Measurement in Education, 1999). La validez ha dejado de ser considerada una propiedad de los instrumentos para convertirse en un cúmulo de evidencia sobre la calidad de las inferencias que se pueden realizar en el proceso de medición. A la luz de este marco integrador se considera el estudio de la estructura cognitiva representada en **Q**. Desde una perspectiva psicométrico-cognitiva, partiendo de la fase inicial de diseño de los ítems, se comienza un programa general de validación del constructo y los procedimientos que se han propuesto para revisar la estructura cognitiva

constituyen piezas que aportan evidencia desde diferentes puntos de vista.

Comúnmente se ha evaluado la calidad de la matriz \mathbf{Q} mediante el ajuste del modelo componencial con el que se está llevando a cabo el proceso diagnóstico. En la mayoría de los modelos basados en la TRI se utiliza la prueba de razón de verosimilitud para establecer la bondad de ajuste del modelo, siendo el desajuste un indicador de que la estructura cognitiva resumida en \mathbf{Q} no es apropiada o es incorrecta. Un ejemplo de dicho enfoque se puede encontrar en Embretson (1984) quien prueba la calidad de diversos modelos cognitivos para un mismo grupo de tareas mediante su ajuste al modelo general de rasgo latente (GLTM).

Otra aproximación frecuente para la evaluación de la calidad de \mathbf{Q} , utilizada especialmente en los modelos basados en clases latentes, es el estudio de tasas de clasificación. Un ejemplo se encuentra en Tatsuoka (1995) quien afirma que un alto porcentaje de examinados sin ser clasificados en los estados de conocimiento mediante el modelo RSM podría indicar una mala especificación de la matriz \mathbf{Q} . Una tercera forma de evaluar la validez de \mathbf{Q} , a través del modelamiento psicométrico, tiene que ver con la estimación parámetros: Rupp y Templin (2008) han encontrado, mediante estudios de simulación, que se produce sobreestimación de los parámetros de desliz (s) del modelo DINA cuando se eliminan atributos y sobreestimación de los parámetros de adivinación (g) cuando se añaden atributos en la matriz \mathbf{Q} . Los efectos encontrados son locales en el sentido en que la sobreestimación de los parámetros de un ítem mal especificado no altera la precisión de la estimación de los parámetros de otro ítem, de ahí que valores elevados en los parámetros s y g puedan utilizarse como índices locales de errores de especificación de la estructura cognitiva. Los autores también reportan altas tasas de clasificación errónea para examinados cuyo patrón de habilidad se corresponde con combinaciones de atributos que han sido completamente eliminadas de la matriz \mathbf{Q} . En otros modelos como el unificado (Dibello, Stout y Roussos, 1995) se ha incluido un parámetro residual que recoge la habilidad no contemplada en los atributos definidos en la matriz \mathbf{Q} y que puede ser usado como indicador del grado en que la estructura especificada en \mathbf{Q} explica el desempeño de los sujetos en el test.

Aunque los estudios suelen apuntar a que la falta de ajuste de los modelos componenciales es causada por errores de especificación en la estructura cognitiva el desajuste puede tener diversas causas como una mala elección del modelo (por ejemplo, cuando los

atributos son dependientes y se elige un modelo conjuntivo), al incumplimiento de los supuestos del modelo o a una pobre estimación de los parámetros; por este motivo también se han propuesto algunos métodos de validación independientes de los modelos que pueden usarse en conjunción con éstos y que otorgan nuevas piezas de evidencia dentro un marco integrado de la validez.

Medina-Díaz (1993) propuso emplear la técnica de asignación cuadrática para comparar la similitud entre las relaciones estructurales derivadas de la matriz \mathbf{Q} y los datos observados. Una limitación de este procedimiento es que no solamente requiere los resultados de la ejecución de los ítems sino que también necesita información del desempeño en cada uno de los procesos involucrados. Corter (1995) planteó la utilización del análisis clúster de forma exploratoria, cuando no se dispone de una teoría previa sobre los atributos involucrados en la solución de las tareas, o confirmatoria, para dar significado adicional a la caracterización de los ítems mediante atributos. Según esta propuesta, las correlaciones entre los ítems son consideradas medidas de proximidad de naturaleza discreta, por tanto una correlación alta indicará que dos ítems requieren el mismo conocimiento y consecuentemente los mismos atributos. Una dificultad de este procedimiento es que los clústeres pueden formarse por otras características como contenidos comunes o posición en el test y no solo por compartir atributos.

Dimitrov y Raykov (2003) han propuesto un procedimiento basado en el modelamiento por ecuaciones estructurales (SEM). Los autores definen una matriz de subordinación (\mathbf{S}_{ij}), cuyos elementos asumen un valor de 1 cuando un ítem i está subordinado al j , es decir, si el grupo de atributos requeridos por el ítem i es un subconjunto propio de los atributos requeridos por el ítem j . Las relaciones de subordinación así definidas se representan luego en un modelo SEM que permite evaluar su calidad. Adicionalmente se propone un mecanismo de triangulación mediante el cual se verifica que la proporción de acierto del ítem i sea mayor que la del j , ya que al estar subordinado a él requerirá menos procesos y consecuentemente se espera que sea más fácil. Si no se verifica lo anterior o si el modelo SEM presenta desajuste y saturaciones no significativas es porque la subordinación puede ser incorrecta y por tanto la estructura de \mathbf{Q} tendrá que ser revisada. Una desventaja de este procedimiento es que requiere el cumplimiento de los supuestos del SEM (tamaño grande de muestra, identificación del modelo y normalidad multivariante).

1.6.1. Método de la distancia minimocuadrática (LSDM)

Recientemente, Dimitrov (2007) ha propuesto el método LSDM para el análisis de atributos y validación de estructuras cognitivas en ítems binarios. Este método utiliza los parámetros de los ítems estimados con alguno de los modelos de la TRI para encontrar la probabilidad de dominio de los atributos cognitivos en niveles fijos de habilidad, por lo general, expresados en la escala logística. El método LSDM es conjuntivo y no compensatorio pues parte del supuesto de que la probabilidad de resolver correctamente un ítem es igual al producto de las probabilidades de dominio de los atributos requeridos para resolverlo; en otras palabras, el dominio de un atributo no compensa la carencia de otros componentes.

La idea general del LSDM es que, conociendo los parámetros de los ítems, se determina la probabilidad de acierto de cada ítem (P_{ij}) en niveles fijos de habilidad. Dado que se ha especificado también la matriz \mathbf{Q} se tiene un sistema de ecuaciones $\mathbf{L} = \mathbf{QX}$ donde \mathbf{X} es la incógnita, \mathbf{L} es el logaritmo neperiano de P_{ij} y \mathbf{Q} es la matriz de especificación cognitiva. El sistema se resuelve encontrando el valor de \mathbf{X} que minimiza la norma del vector $\|\mathbf{QX} - \mathbf{L}\|$. Los valores de la solución, $\exp(\mathbf{X})$, corresponden a la estimación de las probabilidades de dominio de los atributos en los niveles fijos de habilidad, cuya expresión gráfica se denomina Curva de Probabilidad de los Atributos (CPA).

La multiplicación de las probabilidades de dominio de los atributos estimadas con el método, de acuerdo a los requerimientos especificados en \mathbf{Q} , permite obtener para cada ítem una curva recuperada que se puede comparar con la CCI obtenida mediante la TRI. Dicha comparación constituye un criterio para evaluar la calidad de los vínculos entre ítems y atributos que han sido establecidos en la matriz \mathbf{Q} pues un ítem mal recuperado constituye evidencia de errores en la especificación de la estructura cognitiva.

El método LSDM puede ser utilizado en combinación con cualquier modelo compo-nencial aportando evidencia en el proceso integral de validación de la estructura cognitiva. Es una herramienta que puede ser empleada antes del modelamiento para verificar que tan apropiada es la matriz \mathbf{Q} y valorar re-especificaciones de la misma dentro el marco de la teoría en la que se ha desarrollado. Para usar el método no se necesita disponer de las respuestas de los examinados, basta con tener los parámetros estimados con alguno

de los modelos TRI que ajuste a los datos. Los parámetros pueden provenir, por ejemplo, de estudios previos, reportes técnicos o publicaciones, sin necesidad de recolectar datos por lo que el LSDM permite ahorrar tiempo y recursos. El método puede ser aplicado en diferentes contextos y tareas sin requerir largos procesos de recolección de datos como en el caso de la técnica de asignación cuadrática propuesta por Medina-Díaz (1993), ni el cumplimiento de supuestos estrictos como en el procedimiento SEM (Dimitrov y Raykov, 2003).

Una de las características más importantes del método es su parsimonia pues no requiere estimación de parámetros mediante procesos iterativos por lo cual, en principio, no tiene límites en cuanto al número de ítems o atributos que se pueden analizar, tampoco requiere la utilización de software especializado. Los resultados obtenidos con el LSDM son fáciles de transmitir y comprender ya que las curvas de probabilidad de los atributos tienen una interpretación muy similar a la de las CCIIs en la TRI pues están en una escala logística.

Además de aportar evidencia para la validez de estructuras cognitivas, el LSDM tiene otras aplicaciones como: a) comparar las probabilidades de dominio de los atributos entre grupos definidos por características de interés como hombres y mujeres, estudiantes provenientes de diferentes tipos de escuelas, e.t.c, b) puede usarse como herramienta para validar los resultados de otros modelos componenciales como el LLTM o el DINA, c) permite analizar características de los atributos como monotonidad, dificultad relativa y discriminación, d) permite realizar una retroalimentación más detallada a los profesores cuando se utiliza en el contexto de la enseñanza-aprendizaje.

Una aplicación potencial de este método, que no ha sido desarrollada por Dimitrov (2007), es que se puede utilizar para clasificar a los examinados en estados de conocimiento utilizando para ello los parámetros de habilidad estimados con la TRI y las curvas de probabilidad de los atributos. El último estudio de la presente tesis explora con profundidad dicha aplicación potencial.

1.7. Problema y objetivos

1.7.1. Estructura de la tesis

Los modelos componenciales y los procedimientos de validación de la estructura de \mathbf{Q} proporcionan evidencia complementaria dentro de un concepto integral de validez. Dentro de las técnicas propuestas hasta el momento, el LSDM constituye una aproximación prometedora por sus múltiples aplicaciones y potencialidades. Sin embargo, por ser un método recientemente formulado no se ha profundizado en sus propiedades técnicas, y son pocos los estudios empíricos que han aplicado dicha metodología. Es por este motivo que el objetivo general de la presente tesis es investigar las propiedades del LSDM e ilustrar el método con algunas aplicaciones empíricas. Además de la Introducción (Capítulo 1), el Marco Teórico (Capítulo 2) y la Discusión General (Capítulo 7), la tesis contiene tres partes: en el Capítulo 3 se presenta un estudio de simulación en el que se analiza la distribución del estadístico MAD y se determinan puntos de corte para evaluar la recuperación de las CCI's. En los Capítulos 4 y 5 se presentan dos estudios de simulación sobre los efectos de la introducción de errores de especificación en la matriz \mathbf{Q} , uno de dichos trabajos es de carácter descriptivo (Capítulo 4) y el otro inferencial (Capítulo 5). Finalmente, en el Capítulo 6 se presentan aplicaciones del método LSDM. Cada capítulo ha sido escrito como un trabajo individual, con lo cual, el texto de algunos capítulos puede coincidir parcialmente. La bibliografía consultada se expone al final del documento (Capítulo 9). La tesis también incluye un apartado de Anexos (Capítulo 8) en el que se proporciona información adicional sobre el diseño y algunos resultados que no se han incluido en el cuerpo del trabajo.

1.7.2. Estudios de simulación

Uno de los criterios fundamentales para la validación de estructuras cognitivas con el LSDM es la correcta recuperación de las CCI's, sin embargo, ni el autor del método ni otros investigadores han realizado estudios de simulación en los que se analicen las propiedades del estadístico MAD y su comportamiento en diversas condiciones, por lo tanto, se desconoce si este criterio es igualmente eficaz en diversas condiciones, por ejemplo, manipulando la cantidad de atributos que contiene la matriz de especificación cognitiva

y el número de ítems que componen el test. También interesa conocer cómo influye la dificultad de los componentes en la estimación de este estadístico. Tampoco se cuenta con un punto de corte para la clasificación de la bondad de ajuste de los ítems que se pueda generalizar y usar con confianza, pues la regla de decisión que propone Dimitrov (2007) para el MAD se basa en resultados de aplicaciones reales que tienen un bajo grado de generalización.

Para responder a estas cuestiones se ha decidido realizar un primer estudio de simulación que se titula: *Análisis de la recuperación de las CCI's con el LSDM* (Capítulo 3), cuyo objetivo es generar y analizar las distribuciones del estadístico MAD según las características de la matriz \mathbf{Q} (cantidad y dificultad de los atributos y número de ítems en el test) y generalizar puntos de corte para que dicho estadístico pueda ser usado en la valoración de la recuperación de las CCI's en situaciones aplicadas. Para cumplir el objetivo se ha diseñado un estudio de simulación con nueve condiciones obtenidas de la manipulación de tres variables independientes (número de ítems y de atributos y dificultad relativa de los mismos). Se han generado, para cada condición, matrices \mathbf{Q} ficticias, se han simulado las respuestas de los sujetos de acuerdo a la estructura de dichas matrices \mathbf{Q} , se han calibrado los datos con un modelo de TRI y finalmente se ha empleado el LSDM. Este procedimiento se ha replicado 100 veces para trazar la distribución de los valores MAD. Los resultados incluyen el análisis de la distribución de los valores MAD y la obtención de puntos de corte en cada condición.

Por otra parte, el método LSDM se ha propuesto como una herramienta para el análisis de atributos y validación de estructuras cognitivas que puede ser utilizado en conjunción con cualquier modelo de diagnóstico cognitivo, entonces, una de las preguntas primordiales que se pueden plantear es si este modelo tiene la capacidad para detectar especificaciones incorrectas en la matriz \mathbf{Q} y, en caso de ser así, qué tipo de errores se pueden detectar con esta herramienta. Una cuestión derivada de la anterior es cuáles de los tres criterios planteados por Dimitrov (2007) para la validación de estructuras cognitivas con el LSDM se ven afectados cuando existen errores en la especificación cognitiva y en qué medida. Específicamente interesa estudiarlos en cuanto a su capacidad (sensibilidad y especificidad) en la detección de diversas fuentes de error y su comportamiento en diferentes condiciones como tamaño del test, el número de atributos de la matriz \mathbf{Q} y la dificultad de los atri-

butos que se encuentran incorrectamente especificados. Para responder a los problemas planteados anteriormente se han diseñado dos estudios de simulación, uno de ellos bajo un enfoque descriptivo y el otro con un interés inferencial.

El segundo trabajo de simulación de la presente tesis se titula: *Descripción de los efectos de errores de especificación en la estructura cognitiva* (Capítulo 4). El objetivo de este trabajo es analizar de forma descriptiva los efectos de la introducción de cuatro tipos de error de especificación en la matriz \mathbf{Q} sobre los criterios de validación con el LSDM y en la recuperación de las CPAs. Específicamente se han evaluado los efectos de las siguientes fuentes de error: a) modificación en los vínculos concretos entre atributos e ítems, b) eliminación de atributos requeridos y adición de atributos no contemplados en la estructura original, c) introducción de relaciones de dependencia secuencial y de incompatibilidades entre atributos y d) heterogeneidad y falta de monotonicidad en los atributos. En el diseño de esta investigación se ha partido de una matriz \mathbf{Q} ficticia, de 4 atributos y 15 ítems, que ha servido de base para la introducción controlada de los errores de especificación. En este estudio se han manipulado tres variables independientes: la densidad de \mathbf{Q} , la dificultad de los atributos y los tipos de cambio introducidos de acuerdo a los objetivos anteriormente mencionados. En los resultados se describen los efectos de los diversos tipos de error sobre la estimación y forma de las CPAs y sobre la recuperación de las CCI.

El tercer estudio de simulación que compone la tesis se titula: *Sensibilidad del LSDM para la detección de errores en \mathbf{Q}* (Capítulo 5) y su objetivo es analizar las propiedades estadísticas de los criterios MAD y LSD; específicamente se analiza su capacidad para detectar los diferentes tipos de errores de especificación reflejados en la matriz \mathbf{Q} . Para cumplir con este objetivo se han generado 4 matrices \mathbf{Q} ficticias y se han introducido en ellas 14 tipos diferentes de errores de especificación. En segundo lugar se ha simulado, para cada condición, matrices de probabilidad de acierto a los ítems (P_{ij}) basadas en la ecuación básica del LSDM y se han obtenido, mediante un procedimiento de aleatorización, 200 re-muestras de cada matriz P_{ij} . En tercer lugar se ha aplicado el LSDM usando como entrada cada réplica de la matriz de probabilidades P_{ij} y se ha obtenido la distribución de los criterios de interés (MAD y LSD) para cada \mathbf{Q} . Finalmente se ha evaluado la sensibilidad y especificidad de los criterios del LSDM para la detección de errores de

especificación introducidos en la estructura cognitiva, mediante la metodología de curvas ROC (en inglés Received Operating Characteristic; Fawcett, 2006). En los resultados se presenta el promedio (en las 200 réplicas) de los valores LSD y MAD asociados a una Fracción de Falsos Positivos (FFP) de 0.05, el promedio de la Fracción de Verdaderos Positivos (FVP), el promedio del área parcial bajo la curva estandarizada para valores FFP entre 0 y 0.05, y los intervalos de confianza para la media y desviación estándar de todos los estadísticos anteriormente mencionados.

1.7.3. Estudio de aplicación con datos reales

El conocimiento sobre estructuras cognitivas que subyacen al desempeño en tareas aritméticas puede ayudar a los profesores de matemáticas y a los desarrolladores de pruebas a construir ítems con determinadas características tanto a nivel de la exigencia cognitiva como para conseguir propiedades de medida deseables. Un entendimiento más profundo de los atributos de pensamiento, aprendizaje y habilidades involucradas en la solución de los test de Matemáticas tiene muchas ventajas; entre ellas se puede contar un mejor diagnóstico de las fortalezas y dificultades de los alumnos y el desarrollo de estrategias de enseñanza que se enfoquen en procesos cognitivos específicos. Es por este motivo que el objetivo general de esta parte de la tesis es ilustrar la utilidad del método LSDM para validar y analizar los atributos cognitivos propuestos para resolver tareas de adición y sustracción con números enteros. También se pretende explorar la potencialidad del método como herramienta para el diagnóstico cognitivo, comparar los resultados de este método con otros modelos componenciales y utilizar los resultados de los trabajos de simulación previos en un estudio de aplicación con datos reales.

Los estudios de aplicación se han agrupado en un trabajo titulado: *Aplicaciones del LSDM* (Capítulo 6) que tiene varios objetivos específicos. El primer objetivo ha sido diseñar un test de aritmética basado en una estructura teórica de seis atributos cognitivos requeridos para resolver tareas de adición y sustracción entre números enteros; se ha buscado incorporar principios cognitivos en la construcción de las tareas con el interés de conseguir una base apropiada para el diagnóstico cognitivo. Para cumplir dicha meta se ha estudiado profundamente el constructo de habilidad aritmética mediante un análisis sistemático de los procesos, estrategias y atributos mentales que utilizan los estudiantes

para resolver este tipo de operaciones. Se han empleado diversas metodologías para determinar la estructura cognitiva subyacente al test de aritmética entre las que se pueden contar: entrevistas con los profesores de Matemáticas, estudio del currículo y de libros de texto de séptimo grado (curso objetivo), entrevistas y análisis de protocolos verbales con alumnos de séptimo grado de las edades comprendidas entre 12 y 13 años (población objetivo), estudio de la teoría cognitiva que se ha desarrollado hasta el momento y un estudio piloto previo en el que se utilizó la misma estructura cognitiva con un formato de ítems y una muestra diferente. El instrumento se compone de 39 ítems de elección múltiple, de los cuales 13 preguntan por la obtención directa de la respuesta y 26 por los pasos o componentes requeridos para solucionar cada uno de los 13 ítems compuestos. Estos últimos ítems se han denominado ítems individuales y se han diseñado con el objetivo de tener información observada sobre las probabilidades de dominio de los seis atributos hipotetizados.

El segundo objetivo ha sido aplicar el LSDM para acopiar evidencia sobre la validez de la estructura cognitiva propuesta y proponer tres nuevos criterios para estudiar la recuperación de las CCIs. Para llevarlo a cabo se ha aplicado el test de aritmética a una muestra de 382 alumnos de colegios públicos en la ciudad de Bogotá-Colombia. Los datos se han calibrado con el modelo TRI de Rasch tras hacer un análisis de la dimensionalidad y de las propiedades de los ítems mediante la Teoría Clásica de los Test (TCT). Se ha estudiado el ajuste del modelo de TRI a los datos y finalmente se ha ejecutado el método LSDM. Para acopiar evidencia sobre la validez de la estructura propuesta se han usado las propiedades de los criterios, evaluadas en los estudios previos de simulación y se han propuesto tres nuevos estadísticos: los valores MAD ponderados (WMAD) en los que se usa como peso los residuales de la minimización de la norma tras ejecutar el LSDM, la raíz de la diferencia media al cuadrado (RMSD) y la construcción de unos límites para evaluar la calidad de la recuperación de las CCIs basados en el error estándar del parámetro de dificultad estimado con el modelo de Rasch. El último procedimiento solo se ha podido aplicar en el estudio con datos reales pues el error de los parámetros en los estudios simulados era tan bajo que los límites carecían de utilidad.

El tercer objetivo del Capítulo 6 ha sido realizar una validación cruzada de los resultados obtenidos con el LSDM por medio del análisis de los atributos propuestos con otros

modelos componenciales (LLTM y DINA). Para llevarlo a cabo se ha evaluado el ajuste de los dos modelos componenciales y se han usado los parámetros básicos estimados con el LLTM y las prevalencias de los atributos estimadas mediante DINA para establecer un ordenamiento de los atributos en cuanto a su dificultad relativa y ejecutar así comparaciones con el ordenamiento producido por el modelo LSDM y con las proporciones de acierto observadas en los ítems individuales.

Finalmente se han explorado las potencialidades del LSDM como herramienta de diagnóstico cognitivo y se ha estudiado la eficacia de un procedimiento propuesto para tal fin. Para cumplir dicha meta se ha propuesto una metodología, basada en las CPAs estimadas con el LSDM, para clasificar los sujetos en estados de conocimiento. Posteriormente, para valorar dicho procedimiento, se han comparado las probabilidades de dominio de los atributos estimadas con el LSDM y DINA con la proporción de acierto observada en los ítems individuales.

Capítulo 2

Marco Teórico

Este capítulo incluye las bases teóricas en las cuales se sustenta la presente tesis. En las secciones uno y dos se presentan algunos de los principales modelos componenciales que han sido propuestos hasta el momento, se ha decidido dividirlos en dos grupos según sus bases: por una parte se exponen los modelos derivados de la TRI y por la otra los denominados Modelos de Diagnóstico Cognitivo (MDC) . En la sección tres se explica el método de las distancias mínimo-cuadráticas (Dimitrov, 2007) que es el foco de la presente tesis. En la sección cuatro se exponen las características, ventajas y limitaciones de otros métodos para la validación de estructuras cognitivas. En la quinta sección se presentan las implicaciones del enfoque psicométrico-cognitivo en el diseño de pruebas, en el concepto y proceso de validez, en la caracterización de los atributos cognitivos y en la Psicometría en general y finalmente, en la sección seis, se examinan los modelos teóricos cognitivos sobre la habilidad aritmética con números enteros en los que se ha sustentado la estructura propuesta para los estudios empíricos de la presente tesis.

2.1. Modelos basados en la TRI

Los modelos componenciales que se ha decidido incluir en el primer grupo se basan en la TRI. Estos modelos combinan un modelo matemático con un modelo latente de TRI como el modelo de Rasch. El modelo matemático vincula los ítems con atributos teóricos mientras que el modelo TRI especifica cómo combinar parámetros de los examinados y de los ítems en verosimilitudes de respuesta específicas.

2.1.1. Modelos precursores

2.1.1.1. Modelo de Suppes

El modelo probabilístico propuesto por Suppes (1969) es el punto de partida para el desarrollo de muchos modelos probabilísticos componenciales. La ecuación de este modelo es:

$$P_{ij} = \prod_{k=1}^K (1 - \varepsilon_k)^{f_{jk}} \quad (2.1)$$

Donde,

P_{ij} es la probabilidad de que el examinado i resuelva el ítem j .

ε_k es la probabilidad de que el componente k no sea ejecutado correctamente.

f_{jk} es la frecuencia con la cual el componente k es requerido para resolver el ítem j .

La probabilidad ε_k representa en este modelo una tasa de error del componente; sin embargo es la misma para todos los examinados y esto constituye una limitación del modelo que es superada por los que se presentan a continuación.

2.1.2. Modelos Unidimensionales

Este grupo de modelos tiene en común que en ellos se estima un único parámetro de habilidad para cada examinado, es decir, no se considera que la persona tenga un nivel de habilidad para cada componente sino que se asume un nivel global de habilidad para el examinado, por lo tanto estos modelos son de carácter unidimensional.

2.1.2.1. Modelo LLTM

El modelo Lineal Logístico de Rasgo Latente (LLTM), propuesto por Fischer (1973), se define formalmente como una extensión del modelo de Rasch en la que se establecen restricciones lineales sobre sus parámetros. La ecuación del modelo es:

$$P_{ij} = \frac{\exp \left[\theta_i - \sum_{k=1}^K w_{jk} \alpha_k + c \right]}{1 + \exp \left[\theta_i - \sum_{k=1}^K w_{jk} \alpha_k + c \right]} \quad (2.2)$$

Donde,

P_{ij} es la probabilidad de que el examinado i resuelva la tarea o ítem j .

θ_i es la habilidad del examinado i .

α_k son los llamados *parámetros básicos* que evalúan la contribución del componente k a la dificultad del ítem.

w_{jk} es el *peso* del parámetro básico k para el ítem j .

c es una constante de normalización.

Como se puede ver en la ecuación (2.2), el LLTM descompone el parámetro de dificultad en sumas ponderadas de los parámetros básicos y se añade una constante de normalización c . El propósito de este modelo es predecir la dificultad de los ítems o tareas del test a partir de un grupo de componentes cognitivos definidos a priori y expresados en una matriz de pesos \mathbf{W} cuyos elementos son $w_{jk} = 1$ si el ítem j requiere el componente k para su correcta resolución y $w_{jk} = 0$ en otro caso. Nótese que \mathbf{W} es la misma matriz de especificación que posteriormente Tatsuo (1983) denomina matriz \mathbf{Q} . Los valores de α_k se estiman a partir de \mathbf{W} y de las respuestas de los sujetos mediante máxima verosimilitud condicional.

La aplicación correcta del modelo LLTM requiere tres tipos de contrastes:

1. Probar que el modelo de Rasch ajuste correctamente a los datos.
2. Evaluar el ajuste del modelo LLTM. Esto se suele hacer mediante la prueba de Razón de Verosimilitud (RV; Fischer y Ponocny-Seliger, 1998). Por medio de dicho test se evalúan las restricciones lineales de la ecuación (2.2), en concreto, se contrasta la hipótesis nula de que los parámetros obtenidos tras la aplicación del LLTM reproducen adecuadamente los parámetros del modelo de Rasch. El estadístico RV presenta una distribución chi-cuadrado (χ^2) con grados de libertad iguales a la diferencia en el número de parámetros independientes entre los dos modelos (LLTM y Rasch). La prueba RV puede ser complementada con la correlación entre las dificultades estimadas con los dos modelos y su representación gráfica. En las aplicaciones empíricas es frecuente que la prueba RV sea significativa indicando desajuste del LLTM aunque las correlaciones entre los parámetros indiquen una adecuada descomposición de la dificultad de los ítems por parte del modelo. Es por eso que se recomienda basarse más en las correlaciones ya que la RV resulta una prueba muy conservadora (Fischer, 1995).
3. Acopiar evidencia sobre la validez de la estructura representada en \mathbf{W} .

Algunos investigadores han estudiado las propiedades del LLTM como su sensibilidad frente a los errores de especificación en la matriz \mathbf{W} (Baker, 1993), la equivalencia de modelos (Bechger, Verstralen y Verhelst, 2002) y las relaciones del LLTM con otros modelos (Fischer y Tanzer, 1994).

Se han realizado también numerosas aplicaciones del LLTM, pues se trata de uno de los modelos componenciales empleados con mayor frecuencia. En el campo de las Matemáticas se ha aplicado en problemas de álgebra (Fischer, 1973, Medina-Díaz, 1993; Dimitrov y Raykov, 2003), operaciones aritméticas básicas (Real, Olea, Ponsoda, Revuelta y Abad, 1999; Romero, Ponsoda y Ximénez, 2008) y operaciones con fracciones y números mixtos (López y Elosua, 2002). También ha sido aplicado en otras áreas como el razonamiento abstracto (Embretson, 1998; Hornke y Habon, 1986; Tanzer, Gittler y Ellis, 1995) y la comprensión lectora (Gorin, 2007; Embretson y Wetzel, 1987).

2.1.2.2. Modelo ALM

El modelo Logístico Atomatón (ALM) propuesto por Spada (1977) expresa la probabilidad de acierto del ítem como producto de las probabilidades de ejecutar correctamente los atributos requeridos para resolverlo; esta última es determinada por un modelo de Rasch con un parámetro de dificultad para cada atributo y un parámetro de habilidad para cada examinado. La ecuación del modelo es:

$$P_{ij} = \prod_{k=1}^K \left[\frac{\exp(\theta_i - \beta_k)}{1 + \exp(\theta_i - \beta_k)} \right]^{f_{jk}} \quad (2.3)$$

Donde,

P_{ij} es la probabilidad de que el examinado i resuelva el ítem j .

θ_i es la habilidad del examinado i .

f_{jk} es la frecuencia con la que el atributo k es requerido para resolver el ítem j .

β_k es el parámetro de dificultad para el atributo k .

Para la estimación de parámetros en el ALM es necesario utilizar las respuestas de los examinados a subtarefas diseñadas en concordancia con la estructura cognitiva. En otras palabras, el modelo requiere información directa sobre la ejecución del examinado en cada atributo. Este modelo es muy similar al Modelo Multicomponencial de Rasgo

Latente (MLTM, Whitely, 1980), que será expuesto más adelante, la diferencia es que el ALM es unidimensional porque supone un único nivel de habilidad por cada examinado mientras que el MLTM es multidimensional asumiendo diferentes niveles de habilidad por cada atributo.

2.1.2.3. Modelo MIRID

En el modelo con restricciones internas en las dificultades de los ítems o modelo MIRID (Butter, De Boeck y Verhelst, 1998) se realiza una descomposición aditiva del parámetro de dificultad de forma similar al LLTM, la diferencia es que en MIRID se estima un parámetro de dificultad para cada atributo y cada ítem total, mientras que en el LLTM se estima únicamente para cada atributo. En MIRID también se incluye un peso para cada uno de estos parámetros, dichos pesos pueden ser interpretados en términos de su influencia global sobre el ítem, teniendo en cuenta la precisión de su estimación (errores estándar).

Este modelo ha sido propuesto para test diseñados con *familias de ítems*, es decir, tareas compuestas por el ítem total y los procesos o atributos necesarios para resolverlo (subtareas). El modelo MIRID estima los parámetros de dificultad para cada sub-tarea y la dificultad total del ítem es una combinación lineal de los parámetros de las subtareas como se puede observar en la siguiente ecuación:

$$P_{ij} = \frac{\exp \left[\theta_i - \sum_{k=1}^K \sigma_k \beta_{jk} - c \right]}{1 + \exp \left[\theta_i - \sum_{k=1}^K \sigma_k \beta_{jk} - c \right]} \quad (2.4)$$

Donde,

P_{ij} es la probabilidad de que el examinado i resuelva la tarea o ítem j .

θ_i es la habilidad del examinado i .

β_{jk} es el parámetro de dificultad para el componente k y el ítem total j .

σ_k es el *peso* del atributo k en la dificultad del ítem total j .

c es una constante de normalización.

Como se puede ver en la ecuación (2.4) MIRID, al igual que LLTM, es un modelo de Rasch con restricciones internas en los parámetros de dificultad; la diferencia es que en MIRID se estiman parámetros de dificultad para cada componente k dentro de cada ítem

j . Como MIRID es un modelo de Rasch restringido, es posible probar el ajuste del modelo utilizando pruebas de RV para comparar el modelo con y sin restricciones. El estadístico RV tiene distribución χ^2 con grados de libertad igual a la diferencia de parámetros entre los dos modelos.

Los parámetros de dificultad β_{jk} , pesos σ_k , y el intercepto c pueden ser estimados por máxima verosimilitud condicional (CML) condicionados a la puntuación total de los examinados. Para obtener valores de partida adecuados, el algoritmo estima en primer lugar las dificultades de las subtareas y del ítem total con un modelo de Rasch no restringido, luego, las dificultades resultantes son usadas como puntos de partida para la estimación bajo el modelo MIRID. Los valores iniciales para los pesos y el intercepto son obtenidos por regresión lineal de las dificultades de los atributos en el modelo de Rasch. Según Butter, De Boeck y Verhelst (1998) en la mayoría de los casos el algoritmo converge en pocos pasos.

Se han realizado algunas aplicaciones y extensiones del modelo MIRID: Smits y De Boeck (2003) lo utilizan para analizar la descomposición del concepto psicológico de culpa y Smits, De Boeck y Verhelst (2003) proporcionan un programa para la estimación de los parámetros mediante CML.

2.1.2.4. Modelo 2PL-CM

El modelo logístico de dos parámetros constreñido (2PLCM), propuesto por Embretson (1999), es una generalización del modelo TRI de dos parámetros que estima pesos para los parámetros de dificultad y discriminación. La ecuación del modelo es:

$$P_{ij} = \frac{\exp \sum_{k=1}^K q_{jk} \tau_k \left[\theta_i - \sum_{k=1}^K q_{jk} \sigma_k \right]}{1 + \exp \sum_{k=1}^K q_{jk} \tau_k \left[\theta_i - \sum_{k=1}^K q_{jk} \sigma_k \right]} \quad (2.5)$$

Donde,

P_{ij} es la probabilidad de que el examinado i resuelva la tarea o ítem j .

θ_i es la habilidad del examinado i .

q_{jk} es el elemento de la matriz \mathbf{Q} para el ítem j y atributo A_k .

σ_k es el *peso* del atributo A_k en la dificultad del ítem j .

τ_k es el *peso* del atributo A_k en la discriminación del ítem j .

Los parámetros del modelo se pueden estimar por medio de máxima verosimilitud conjunta (JML). Al igual que en otros modelos, el ajuste del 2PLCM se puede evaluar mediante pruebas RV ($\sim \chi^2$) con grados de libertad igual a la diferencia de parámetros entre los modelos que se pretende comparar. Embretson (1999a) presenta una aplicación de este modelo a tareas de razonamiento abstracto y evalúa el ajuste del 2PLCM comparándolo con el modelo de Rasch, el modelo logístico de dos parámetros y el modelo LLTM. La autora encuentra que el modelo 2PLCM presenta un ajuste significativamente mejor que el de los otros modelos.

2.1.3. Modelos Multidimensionales

Este grupo de modelos se caracterizan porque en ellos se considera que los examinados poseen un nivel de habilidad diferente para cada atributo o proceso requerido en la resolución correcta de los ítems, esto implica la presencia de múltiples dimensiones.

2.1.3.1. Modelo MLTM

En el modelo multicomponente de rasgo latente (MLTM), propuesto por Whitely (1980), la probabilidad de responder un ítem correctamente corresponde a la probabilidad conjunta de ejecutar satisfactoriamente los atributos requeridos para resolverlo; dicha probabilidad depende de la dificultad del atributo y de la habilidad del examinado en cada atributo. La ecuación del modelo es:

$$P_{ij} = \prod_{k=1}^K \frac{\exp(\theta_{ik} - \beta_{jk})}{1 + \exp(\theta_{ik} - \beta_{jk})} \quad (2.6)$$

Donde,

P_{ij} es la probabilidad de que el examinado i resuelva la tarea o ítem j .

θ_{ik} es la habilidad del examinado i en el atributo A_k .

β_{jk} es el parámetro de dificultad para el componente A_k en el ítem j .

Como puede observarse en la ecuación (2.6), la probabilidad de resolver correctamente el ítem es el producto de la verosimilitud del atributo A_k , donde dicha verosimilitud es una función logística de la habilidad del examinado para el atributo y la dificultad del atributo en el ítem j . Al igual que en el ALM, la relación entre atributos es no compensatoria,

en otras palabras, la baja probabilidad de dominar un atributo no es compensada por el dominio de otros atributos. A diferencia de los modelos presentados anteriormente, este modelo es multidimensional porque asume múltiples habilidades en los examinados (una para cada atributo).

Para estimar los parámetros del modelo se requiere información sobre las respuestas de los examinados en cada subtarea o atributo individual. Con estos datos se estiman los parámetros locales de cada atributo y luego se obtienen las dificultades de los ítems aplicando el modelo de Rasch en cada atributo separadamente. Finalmente, los parámetros se conjugan de acuerdo con la ecuación (2.6) para estimar las probabilidades de acierto en cada ítem (P_{ij}). Las estimaciones de los parámetros del MLTM también pueden ser realizadas maximizando la verosimilitud de las respuestas conjuntas (a los atributos individuales y a los ítems totales) sobre los examinados, este procedimiento fue propuesto por Maris (1999).

2.1.3.2. Modelo GLTM

El modelo de rasgo latente general (GLTM; Embretson, 1984) combina el LLTM y el MLTM por medio de la siguiente ecuación:

$$P_{ij} = (a - g) \left[\prod_{k=1}^K \frac{\exp(\theta_{ik} - \sum_{k=1}^K w_{jk} \alpha_k + c)}{1 + \exp(\theta_{ik} - \sum_{k=1}^K w_{jk} \alpha_k + c)} \right] \quad (2.7)$$

Donde,

P_{ij} es la probabilidad de que el examinado i resuelva la tarea o ítem j .

θ_{ik} es la habilidad del examinado i en el atributo A_k .

α_k son los parámetros básicos que evalúan la contribución del componente A_k a la dificultad del ítem j .

w_{jk} es el peso del parámetro básico α_k para el ítem j .

c es una constante de normalización.

a es la probabilidad de respuesta correcta de un ítem cuando todos los atributos requeridos son dominados por el examinado.

g es la probabilidad de respuesta correcta de un ítem cuando al menos un atributo requerido no es dominado por el examinado.

Al igual que en el MLTM, en el GLTM la probabilidad de respuesta correcta de un ítem (P_{ij}) es el producto de las verosimilitudes de los atributos, la diferencia es que el GLTM incluye dos parámetros nuevos: el parámetro a , que representa la probabilidad de responder correctamente a un ítem cuando el examinado domina todos los atributos requeridos para resolverlo y el parámetro g , que corresponde a la probabilidad de respuesta correcta en un ítem cuando el examinado carece de dominio de alguno de los atributos implicados en la resolución de dicho ítem, por lo tanto, dicho parámetro representa soluciones alternativas como adivinación al azar o copia.

El GLTM incluye componentes de habilidad para cada atributo (por ello es un modelo multidimensional) y descompone la dificultad del ítem en sumas ponderadas de los parámetros básicos como en el LLTM. Al igual que en el MLTM los parámetros se estiman por máxima verosimilitud conjunta (JML).

El MLTM y el GLTM han sido aplicados en diversas temáticas como en la medición de analogías verbales (Embretson, 1984), comprensión lectora (Embretson y Wetzel, 1987), inteligencia (Embretson, 1996) y de tareas complejas (Embretson y Yang, 2006).

2.1.4. Modelos para la medición de estrategias

Estos modelos buscan analizar las múltiples estrategias para resolver los ítems de un test en el contexto de los modelos componenciales basados en la TRI.

2.1.4.1. Modelo de Mislevy y Verhest (1990)

En este modelo se asume que los ítems del test se responden por diferentes estrategias de carácter exhaustivo y mutuamente excluyente, aunque no todas las estrategias producen respuestas correctas ni son usadas por todos los examinados. Los supuestos del modelo son: a) cada examinado aplica la misma estrategia en todos los ítems del grupo, b) las respuestas de los sujetos son observables pero la estrategia empleada no lo es, c) las respuestas de los examinados que siguen una estrategia conforman un modelo TRI conocido, d) existe teoría sustantiva sobre los atributos cognitivos que conforman el test. La ecuación del modelo es:

$$P(X_{ij} = 1) = \prod_k \left[\prod_j (f_k(\theta_{ik}, \beta_{jk}))^{x_{ij}} (1 - f_k(\theta_{ik}, \beta_{jk}))^{1-x_{ij}} \right]^{\Phi_{ik}} \quad (2.8)$$

Donde,

X_{ij} es la respuesta del examinado i al ítem j .

θ_{ik} es la habilidad del examinado i en la estrategia k .

β_{jk} es la dificultad del ítem j en la estrategia k .

Φ_{ik} es 1 si el examinado usa la estrategia k y 0 en otro caso.

$f_k(\theta_{ik}, \beta_{jk})$ es la función del modelo TRI que ajusta a los datos.

Este modelo también puede ser empleado descomponiendo los parámetros de dificultad de las estrategias (β_{jk}) en *parámetros básicos* al igual que en el modelo LLTM, como se refleja en la siguiente ecuación:

$$\beta_{jk} = \sum q_{jkm} \alpha_{km} \quad (2.9)$$

En este caso el parámetro de dificultad del ítem j , resuelto mediante la estrategia k , es una suma ponderada de los elementos de α_{km} , que corresponde al vector de parámetros básicos asociados con la estrategia k . Los pesos (elementos q_{jkm}) indican el grado en el cual cada atributo, que es relevante en la estrategia k , está presente en el ítem j . Mislevy y Verhelst (1990) aplican este modelo de estrategias en un test de visualización espacial.

2.1.4.2. Modelo MLTM-E

El modelo multicomponente de rasgo latente para la medición de estrategias, propuesto por Embretson en 1985, es una extensión del MLTM que incluye combinación de atributos y uso de diversas estrategias. El modelo incluye la definición de atributos que conforman la estrategia y la especificación del orden en la ejecución de las estrategias. Por ejemplo, si la solución del ítem requiere dos estrategias y en la primera de ellas (k_1) se utilizan dos atributos, la probabilidad de respuesta correcta de dicho ítem es:

$$P(X_{ij} = 1|k_1) = a_1 P_{A_1 k_1} P_{A_2 k_1} \quad (2.10)$$

Donde,

a_1 es la probabilidad de aplicación de la estrategia k_1 .

$P_{A_1 k_1}$ y $P_{A_2 k_1}$ son las probabilidades de resolver correctamente los atributos A_1 y A_2 de la estrategia k_1 dados por el modelo MLTM.

Suponiendo que en la segunda estrategia (k_2) únicamente se utiliza un atributo, que la probabilidad de aplicación de dicha estrategia es a_2 y que adicionalmente k_2 solo puede ser utilizada si el examinado falla uno o dos atributos de la estrategia anterior (k_1), entonces, la probabilidad de resolver correctamente el ítem j utilizando k_2 está dada por la siguiente expresión:

$$P(X_{ij} = 1|k_2) = a_2 P_{A_1 k_2} (1 - P_{A_1 k_1} P_{A_2 k_1}) \quad (2.11)$$

Es decir, la efectividad de la segunda estrategia por la probabilidad de fallar la primera estrategia.

También es posible construir un modelo matemático para la probabilidad de que el ítem haya sido resuelto por adivinación:

$$P(X_{ij} = 1|g) = g (1 - P_{A_1 k_1} P_{A_2 k_1}) (1 - P_{A_1 k_2}) \quad (2.12)$$

Como se puede ver en la ecuación (2.12), la probabilidad de que el ítem j sea resuelto por azar es la multiplicación de las probabilidades de fallo en ambas estrategias multiplicadas por el parámetro de adivinación g .

Una implicación del MLTM-E es que permite estimar la probabilidad de acierto de cada ítem para un examinado que utiliza determinada estrategia y también determinar la mejor estrategia para un ítem utilizando la suma de las habilidades promedio. La autora presenta una aplicación del modelo en una prueba de analogías verbales y estudia su ajuste por medio de la comparación de las prueba RV entre modelos alternativos. En dicha aplicación Embretson (1985) también muestra como la comparación entre modelos permite conocer las estrategias más importantes para resolver el test.

2.2. Modelos de Diagnóstico Cognitivo

Existe un grupo de modelos psicométrico-cognitivos, llamados Modelos de Diagnóstico Cognitivo (MDC), que basan su desarrollo en ideas de los modelos de clase latente (MCL; Lazarsfeld y Henry, 1968). La ecuación general de los MCL es:

$$P(X_i) = \sum_{c=1}^C \eta_c \prod_{j=1}^J \pi_{jc}^{x_{ij}} (1 - \pi_{jc})^{1-x_{ij}} \quad (2.13)$$

Donde,

η_c es la probabilidad de que un examinado sea miembro de la clase c .

x_{ij} es la respuesta observada de un examinado i en el ítem j .

π_{jc} es la probabilidad de que un examinado de la clase c responda correctamente al ítem j o $P(x_{ij} = 1|c)$

El modelamiento por clases latentes proporciona tres tipos de información: a) la proporción de sujetos en cada clase, b) las probabilidades de responder correctamente cada ítem según la clase a la que se pertenece y c) las probabilidades posteriores de pertenencia a la clase para cada examinado. Existen dos aproximaciones diferentes para identificar clases latentes: el enfoque exploratorio y el confirmatorio. Por lo general los MCL son procedimientos exploratorios en los que se busca identificar y caracterizar clases de examinados de acuerdo a sus patrones de respuesta. El modelo mixto de Rasch (MIRA; Rost, 1990) es un ejemplo de aproximación exploratoria puesto que las clases latentes son identificadas empíricamente con base en patrones de respuesta similares.

La idea general de los MDCs, es identificar clases latentes de examinados basadas en las estructuras de conocimiento que las personas utilizan para resolver los ítems, por lo tanto, los MDC representan restricciones confirmatorias de los MCL. La primera de ellas se refiere al número de clases: cuando se tienen k atributos se definen un total de 2^k clases pues los estados de conocimiento son de carácter binario, ya que expresan el dominio o no dominio de los atributos. Los MDC también incluyen restricciones de igualdad determinadas por la selección del modelo a usar y por la especificación de la matriz \mathbf{Q} que representa la estructura cognitiva. El modelo *rule space* (Tatsuoka, 1985), los modelos DINA y NIDA (Junker y Sijtsma, 2001) y el modelo híbrido de Yamamoto y Gitomer (1993) son algunos de los métodos confirmatorios que clasifican al examinado en clases correspondientes a estados o estructuras de conocimiento. El objetivo general de los MDCs es la clasificación de los examinados en patrones, generalmente binarios, que reflejan su habilidad en cada uno de los atributos cognitivos.

2.2.1. Modelo Rule Space

El modelo *rule space* (RSM; Tatsuoaka, 1983) es un modelo para el diagnóstico cognitivo y la clasificación de examinados propuesto por Tatsuoaka (1983, 1984b, 1985). El modelo analiza los patrones de respuesta de los examinados para clasificarlos en los denominados *estados de conocimiento*, que son patrones binarios en los que se refleja el dominio de los sujetos en los atributos latentes.

Para definir el modelo es necesario seguir los siguientes pasos: en primer lugar, se determinan y definen los atributos cognitivos que subyacen a la ejecución del test, esto se suele realizar mediante varios procedimientos entre los que se puede contar el estudio de la teoría cognitiva subyacente, la opinión de expertos en el área de interés y el estudio sistemático de las tareas cognitivas, entre otros. En segundo lugar, se deben establecer los vínculos entre atributos e ítems. En este paso se define una matriz \mathbf{Q} binaria cuyos elementos, q_{jk} , representan si el atributo k es necesario para resolver el ítem j ($q_{jk} = 1$) o no ($q_{jk} = 0$). Una adecuada especificación de dicha matriz es muy importante para la correcta ejecución de los procedimientos de clasificación que se realizan posteriormente con el RSM.

En tercer lugar, se determinan los estados de conocimiento. Como la ejecución de los examinados en los atributos latentes no es directamente observable, tiene que ser inferida de los patrones de respuesta. Idealmente, un examinado que domine todos los atributos requeridos deberá resolver correctamente todos los ítems, mientras que aquel que solo domine algunos atributos responderá correctamente solo aquellos ítems que requieren los atributos que el sujeto domina y responderá incorrectamente aquellos ítems que requieren atributos que no domina. Siguiendo esta lógica, cada examinado tiene un patrón de respuestas ideal y dicho patrón puede ser expresado en términos de atributos o estados de conocimiento.

A modo de ilustración supóngase una matriz \mathbf{Q} con 3 ítems y 2 atributos (véase Tabla 2.1). Con 3 ítems existen 8 patrones de respuesta posibles (111, 110, 101, 100, 011, 001, 010, 000) y con dos atributos se pueden formar cuatro ($2^2 = 4$) estados de conocimiento o combinaciones de dominio-no dominio (ver Tabla 2.1). Ahora bien, de los 8 patrones de respuesta solo 4 son adecuados para la estructura cognitiva expresada en \mathbf{Q} porque

corresponden a algún estado de conocimiento, estos cuatro patrones de respuesta son llamados por Tatsuoka *patrones de respuesta ideales*. Por ejemplo, un examinado que domine el primer atributo debería idealmente tener un patrón de respuesta '100' pues solo acertará el primer ítem que requiere el dominio de A_1 y fallará los otros dos ítems que requieren el dominio de A_2 y de A_1 y A_2 respectivamente, en otras palabras, al estado de conocimiento '10' le corresponde el patrón de respuesta ideal '100', (señalado con negrita en la Tabla 2.1). Los patrones de respuesta restantes que se alejan en mayor o menor medida de los patrones ideales corresponden a factores no contemplados en la matriz \mathbf{Q} como deslices, azar, copia o especificaciones erróneas de \mathbf{Q} .

Tabla 2.1: Matriz \mathbf{Q} , estados de conocimiento (EC) y patrones de respuesta ideales de cada EC.

\mathbf{Q}			EC		ideales		
	A_1	A_2	A_1	A_2	i1	i2	i3
i1	1	0	1	1	1	1	1
i2	0	1	1	0	1	0	0
i3	1	1	0	1	0	1	0
			0	0	0	0	0

Siguiendo la lógica del procedimiento presentado anteriormente es posible establecer correspondencias entre los patrones de respuesta ideales y los estados de conocimiento (Tatsuoka, 1995). Como el número de estados de conocimiento es 2^k siendo k el número de atributos, en algunas ocasiones el número de estados de conocimiento llega a ser muy grande y su cantidad se incrementa exponencialmente cuantos más atributos se requieran. Sin embargo, no todos los estados de conocimiento son relevantes según los vínculos expresados en la matriz \mathbf{Q} porque en muchas ocasiones existen relaciones entre atributos (por ejemplo uno puede ser requisito de otro) que hacen que varios patrones de atributo se correspondan con un solo patrón de respuestas, esta propiedad es llamada *indeterminación de los estados de conocimiento* (Tatsuoka, 1995). Dicha propiedad permite reducir sustancialmente el número de estados de conocimiento en las aplicaciones prácticas del RSM. Un ejemplo de la capacidad de reducción de los estados de conocimiento mediante la propiedad de indeterminación es reportado por Tatsuoka en 1995: la estructura cognitiva propuesta para la subprueba de matemáticas del test SAT tiene 15 atributos y por lo tanto $2^{15} = 32768$ estados de conocimiento, la autora muestra como dichos estados son

reducidos a 1675 usando la propiedad; adicionalmente Tatsuoka afirma que realizar dicha reducción no afecta el proceso posterior de clasificación de examinados.

El penúltimo paso del RSM es la formulación de un espacio de clasificación. Tatsuoka (1983) y Tatsuoka y Tatsuoka (1987) definen un espacio tridimensional de clasificación cuyas dimensiones son: theta (θ), zeta (ξ) y zeta generalizada (ξs). Theta es una variable continua que corresponde al rasgo latente estimado con el modelo TRI, zeta es una medida de ajuste que evalúa el grado de atipicidad de los patrones de respuesta, cuanto mayor es zeta en valor absoluto menos común o más atípico es el patrón de respuesta (Tatsuoka, 1984a). Las zetas generalizadas evalúan que tan atípicos son los patrones de respuesta en subgrupos de n ítems.

En el espacio multidimensional los puntos representan los patrones de respuesta y los patrones ideales que han sido previamente determinados. Como la ejecución del examinado en el test presenta fluctuaciones (aplicación errónea de las reglas, errores conceptuales, deslices, azar, copia...) es raro encontrar un emparejamiento perfecto entre los patrones de respuesta observados y los ideales. Los patrones de respuesta, desviados en mayor o menor medida de los patrones ideales, se representan en el espacio de clasificación como puntos alrededor del centroide, que es el patrón de respuesta ideal más parecido a ellos.

La última etapa del RSM es la clasificación de examinados en estados de conocimiento. Para realizar dicha clasificación se agrupan los patrones de respuesta observados de acuerdo con su cercanía a los patrones ideales previamente establecidos usando la distancia de Mahalanobis. Una vez se ha adjudicado al examinado en un patrón de respuesta ideal se clasifica en el estado de conocimiento que le corresponde. Siguiendo este procedimiento se genera, para cada examinado, una lista de probabilidades de dominio en cada atributo; dicha información es muy útil para propósitos diagnósticos pues permite detectar fortalezas y dificultades individuales en los procesos implicados en la resolución de las tareas.

Tatsuoka ha estudiado las propiedades de el RSM mediante simulaciones en las que se valoran los errores de clasificación y las probabilidades posteriores de pertenecer a una clase basadas en las probabilidades de que los examinados realmente provengan del patrón ideal en el cual han sido clasificados. También se han aplicado reglas de decisión bayesianas para minimizar errores en la clasificación de los examinados en los estados de conocimiento (Tatsuoka, 1995).

El RSM ha sido aplicado principalmente en el análisis de procesos de medición educativa a gran escala, por ejemplo en el área de Matemáticas del estudio TIMSS (Tatsuoka, Corter y Tatsuoka, 2004; Xin, Xu y Tatsuoka, 2004; Birenbaum, Tatsuoka y Xin, 2005) y en el subtest cuantitativo de la prueba GRE (Tatsuoka y Boodoo, 2000).

2.2.2. Modelo DINA

El modelo *Deterministic Inputs, Noisy And Gate* (en inglés, DINA), propuesto inicialmente por Macready y Dayton (1977) y adaptado para propósitos de diagnóstico por Junker y Sijtsma (2001), es un modelo no compensatorio y conjuntivo para la ejecución de ítems o tareas. Se ha definido como conjuntivo porque emplea una función en la cual todos los atributos requeridos tienen que dominarse para la correcta ejecución de los ítems, en otras palabras, los atributos son considerados de manera independiente. Su carácter no compensatorio implica que la carencia de dominio de un atributo no es compensada por el dominio en otros atributos.

En el modelo DINA los examinados se dividen en dos grupos para cada ítem: el conjunto de sujetos que poseen todos los atributos requeridos para resolver correctamente el ítem y el grupo de los que carecen de dominio en al menos uno de los atributos requeridos. Dicha división es capturada en una variable latente determinística ξ_{ij} que toma el valor de 1 si el sujeto posee todos los atributos y 0 en el caso contrario. Es posible que un examinado que domina todos los atributos requeridos falle el ítem por desliz o distracción y también puede ocurrir que un examinado que no domina todos los atributos acierte el ítem por adivinación, uso de otras estrategias no contempladas en \mathbf{Q} , copia o cualquier otro mecanismo. El modelo DINA tiene en cuenta este tipo de situaciones y las representa en dos parámetros: el de adivinación (g_j) que corresponde a la probabilidad de acertar el ítem cuando el sujeto no posee todos los atributos requeridos y el de desliz (s_j) que representa la probabilidad de fallar el ítem cuando el examinado domina todos los atributos requeridos. La ecuación del modelo DINA es:

$$P(X_{ij} = 1 | \boldsymbol{\alpha}_i, s_j, g_j) = (1 - s_j)^{\xi_{ij}} g_j^{1 - \xi_{ij}} \quad (2.14)$$

Donde,

X_{ij} es la respuesta del examinado i al ítem j .

$\xi_{ij} = \prod_{k=1}^K \alpha_{ik}^{q_{jk}}$ es una variable latente determinística que indica si el sujeto i posee todos los atributos requeridos por el ítem j . Toma el valor de 1 si el sujeto posee todos los atributos y 0 en el caso contrario, en dicha expresión q_{jk} es el elemento de la matriz \mathbf{Q} para el ítem j y el atributo A_k y α_{ik} son las variables latentes dicotómicas para el examinado i en el atributo A_k que indican dominio o no dominio del atributo.

$$s_j = P(X_{ij} = 0 | \xi_{ij} = 1)$$

$$g_j = P(X_{ij} = 1 | \xi_{ij} = 0)$$

Como se ve en la ecuación (2.14), el modelo DINA también puede ser descrito como un modelo lógico con entradas determinísticas y con salidas probabilísticas (estocásticas) donde el elemento determinístico es capturado por la variable latente que asume el valor de 1 si el examinado domina todos los atributos requeridos y 0 en otro caso. Para cada ítem se estiman los parámetros s_j y g_j y para cada sujeto se estiman k variables latentes (α_{ik}) que conforman el vector de estados de conocimiento de los examinados.

Si el modelo especificado en la matriz \mathbf{Q} es adecuado para los datos y los ítems presentan buen ajuste, los valores de s_j y g_j serán bajos. Por el contrario, si son altos, indicará desajuste del ítem correspondiente o una mala especificación del mismo en la matriz \mathbf{Q} . De la Torre y Douglas (2004) han puntualizado que el parámetro g en este contexto asume una interpretación más general, ya que no solo corresponde al acierto por azar sino también al uso de estrategias alternativas no incluidas en la matriz \mathbf{Q} . A diferencia del RSM, en el modelo DINA el número de atributos no hace que el modelo se vuelva demasiado complejo porque sólo se estiman dos parámetros por ítem independientemente del número de atributos. Esto convierte al modelo DINA en uno de los modelos más parsimoniosos e interpretables entre los MDCs propuestos hasta el momento.

Los parámetros de este modelo se suelen estimar por máxima verosimilitud conjunta (JML) aunque, recientemente, de la Torre y Douglas (2004) han propuesto otras dos técnicas para la estimación de parámetros, por máxima verosimilitud condicional (CML) y mediante cadenas Markov de Montecarlo (MCMC) y han evidenciado mediante estudios empíricos y de simulación que los algoritmos propuestos proporcionan estimaciones fiables

de los mismos. Los algoritmos para la calibración y algunas aplicaciones del modelo DINA con datos reales en tareas de sustracción de fracciones se presentan en de la Torre y Douglas (2004, 2008). Junker y Sijtsma (2001). McReady y Dayton (1977) y Haertel (1989) también proporcionan aplicaciones del modelo aunque no emplean específicamente el término DINA.

2.2.3. Modelo NIDA

El modelo *Noisy Inputs, Deterministic And Gate* (en inglés, NIDA) ha sido introducido por Maris (1999). Al igual que el modelo DINA, el modelo NIDA involucra variables de respuesta latentes determinadas de forma conjuntiva, sin embargo, *noisy inputs* se refiere a la naturaleza estocástica bajo la cual se determinan dichas variables latentes a partir de α_i . Una diferencia importante entre los modelos DINA y NIDA es que en DINA se estiman los parámetros a nivel del ítem mientras que en NIDA se estiman a nivel de los atributos.

En el modelo NIDA se define una variable de respuesta latente, η_{ijk} , que indica si el examinado i aplica correctamente el atributo k requerido por el ítem j . De nuevo se definen las probabilidades s y g al nivel de las variables de respuesta latente: $s = P(\eta_{ijk} = 0 | \alpha_{ik} = 1)$ y $g = P(\eta_{ijk} = 1 | \alpha_{ik} = 0)$. De acuerdo con el modelo, la respuesta a un ítem, X_{ij} , será correcta si todas las respuestas latentes son satisfactorias, $X_{ij} = \prod_{k=1}^K \eta_{ijk}$. La función de respuesta al ítem para este modelo es:

$$P(X_{ij} = 1 | \alpha_i, \mathbf{s}, \mathbf{g}) = \prod_{k=1}^K \left[(1 - s_k)^{\alpha_{ik}} g_k^{1 - \alpha_{ik}} \right]^{q_{jk}} \quad (2.15)$$

Donde,

X_{ij} es la respuesta del examinado i al ítem j .

q_{jk} es el elemento de la matriz \mathbf{Q} para el ítem j y el atributo A_k .

α_{ik} son las variables latentes dicotómicas para el examinado i en el atributo A_k .

La estimación de parámetros del modelo NIDA se realiza por medio de algoritmos MCMC. El modelo presentado aquí, al igual que el expuesto en Junker y Sijtsma (2001) es una simplificación del modelo presentado por Maris (1999) en el cual los parámetros s y g también varían a nivel de los ítems.

2.2.4. Modelo Unificado

Este modelo, propuesto por DiBello, Stout y Roussos (1995), se denomina unificado porque vincula aproximaciones determinísticas y estocásticas: utiliza el elemento determinístico para clasificar a los examinados en estados de conocimiento y tiene en cuenta errores aleatorios como elementos estocásticos. El modelo incluye cuatro posibles fuentes de error para explicar la variabilidad en las respuestas del ítem: la primera es la estrategia usada por el examinado ya que ésta puede ser diferente a la contemplada en la matriz \mathbf{Q} , la segunda es la incorrecta especificación de la matriz \mathbf{Q} , en concreto se contempla que dicha matriz puede estar incompleta al no incluir todos los atributos requeridos por el ítem.

La tercera fuente de error es la denominada *positividad*, que asume un valor alto cuando el examinado que domina un atributo lo utiliza correctamente y el que no lo domina falla los ítems que requieren dicho componente. Valores bajos de positividad implican probabilidades altas de adivinación y desliz (parámetros g y s en los modelos DINA y NIDA). El error aleatorio restante, que no ha sido explicado por otras fuentes de error, pertenece a la última categoría llamada *slips*. Esta categoría incluye errores inesperados, por ejemplo, aquellos causados por la falta de rapidez. El termino *slip* o desliz es empleado de manera diferente que en los modelos DINA y NIDA ya que en el modelo unificado tiene un sentido más general.

Para cuantificar las dos primeras fuentes de error, DiBello, Stout y Roussos (1995) introducen un nuevo parámetro llamado habilidad latente residual (θ), que puede ser unidimensional o multidimensional. En el modelo unificado, el espacio de habilidad latente está constituido por los atributos especificados en la matriz \mathbf{Q} denotados por α_{ik} y por la habilidad latente no representada en \mathbf{Q} contenida en el parámetro θ . Según los autores, este parámetro también tiene el rol de capturar habilidades cognitivas de alto orden por lo que es esperable que si se prescinde de este parámetro la parte discreta del espacio latente no presente buen ajuste. La siguiente fórmula representa la ecuación del modelo con una determinada estrategia representada en \mathbf{Q} :

$$P(X_{ij} = 1 | \alpha, \theta_i) = (1 - p) [d_j s_{\alpha j} P_j(\theta + 2c_j) + (1 - d_j) P(\theta)] \quad (2.16)$$

Donde,

X_{ij} es la respuesta del examinado i al ítem j .

α es el vector de variables latentes dicotómicas (de elementos α_{ik}) para el examinado i y el atributo A_k .

p es la probabilidad de desliz (*slip*).

d_j es la probabilidad de seleccionar la estrategia representada en la matriz \mathbf{Q} para el ítem j .

c_j es el grado en el que \mathbf{Q} contiene los atributos adecuados para resolver el ítem j .

$P_j(\theta)$ es la probabilidad estimada con el modelo TRI con un parámetro de dificultad b_j .

$s_{\alpha j} = \prod_{k \in N(\alpha)} \pi_{kj} \prod_{k \in M(\alpha)} r_{kj}$ es el parámetro de positividad donde,

π_{kj} es la probabilidad de aplicar correctamente el atributo A_k al ítem j dado que $\alpha_{ik} = 1$

r_{kj} es la probabilidad de aplicar correctamente el atributo A_k al ítem j dado que $\alpha_{ik} = 0$

$N(\alpha)$ y $M(\alpha)$ son grupos de los atributos requeridos por el ítem j . Corresponden a atributos dominados y no dominados en el vector α respectivamente.

Como puede verse en la ecuación (2.16) el modelo unificado incluye muchos parámetros para diversos elementos psicométricos. Específicamente el modelo contiene $2k + 3$ parámetros, donde k es el numero de atributos. Un gran inconveniente de este modelo es que no puede ser calibrado porque no está identificado, sin embargo, se ha decidido incluirlo en este marco porque contiene ideas importantes no contempladas en otros modelos ya que los autores han intentado integrar elementos claves de la unión entre Psicometría y Cognición en un solo modelo como son: la clasificación de examinados en estados de conocimiento, la evaluación de la calidad de \mathbf{Q} , el uso de estrategias y procesos cognitivos de orden superior y la posibilidad de modelar los patrones de respuesta erróneos mediante el parámetro de positividad. Adicionalmente, dicho modelo ha sido el punto de partida del que se presentará a continuación.

2.2.5. Modelos RUM - Fusión

Hartz (2002) retomó algunos parámetros del modelo unificado y los combinó para obtener un modelo estadísticamente identificable, el modelo resultante ha sido llamado modelo unificado reparametrizado (RUM) porque busca mantener la propiedad fundamental del unificado al unir elementos psicométricos con el diagnóstico cognitivo. El modelo unificado tiene $2k + 3$ parámetros, mientras que el modelo RUM tiene solo $k + 2$ parámetros por ítem. La ecuación del modelo es:

$$P(X_{ij} = 1 | \alpha, \eta_i) = \pi_j^* \prod_{k=1}^K r_{jk}^{1 - \alpha_{ik} \cdot q_{jk}} P(\theta_i) \quad (2.17)$$

Donde,

X_{ij} es la respuesta del examinado i al ítem j .

α es el vector de variables latentes dicotómicas (de elementos α_{ik}) para el examinado i y el atributo A_k .

q_{jk} es el elemento de la matriz \mathbf{Q} para el ítem j y el atributo A_k .

θ_i es la habilidad latente residual para el examinado i .

$P(\theta_i)$ es la probabilidad estimada con el modelo TRI con un parámetro de dificultad b_j .

$\pi_j^* = \prod_{k=1}^K (1 - s_{jk})^{q_{jk}}$ es la probabilidad máxima de acierto del ítem j .

$r_{jk}^* = \frac{g_{jk}}{1 - s_{jk}}$ es una penalización por cada atributo que no posee el examinado.

s y g en las anteriores expresiones son parámetros de desliz y azar definidos de la misma manera que en los modelos DINA y NIDA.

En la ecuación (2.17) se puede observar que en el RUM se estiman tres parámetros para los ítems (π_j^* , r_{jk}^* y b_j). Cada ítem está caracterizado por un parámetro π_j^* y b_j y k parámetros r_{jk}^* . El parámetro π_j^* refleja la dificultad del ítem j puesto que manifiesta la probabilidad de que un examinado que domina todos los atributos requeridos por el ítem los utilice y aplique correctamente. Este parámetro asume valores entre 0 y 1 y cuanto más alto es su valor estimado, más fácil es el ítem.

El parámetro r_{jk}^* es un indicador inverso de la importancia del atributo A_k en la solución del ítem j , en otras palabras, cuanto más depende la ejecución del ítem del dominio del atributo, más baja es la probabilidad de que un examinado que no domina el

atributo pueda acertar el ítem y consecuentemente más bajo el valor de r_{jk}^* . Cuanto más se acerca dicho parámetro a 0, más discriminativo es el ítem para la habilidad A_k , por el contrario, un valor alto en este parámetro implica que el atributo no es importante para resolver correctamente el ítem j (Hartz, 2002).

$P(\theta_i)$ es un elemento importante en el modelo puesto que cuantifica el grado en el cual la función de respuesta al ítem está explicada por atributos que no han sido especificados en la matriz \mathbf{Q} . Como consecuencia de lo anterior, el parámetro b_j es un indicador de qué tan completa es la especificación del ítem: un valor alto implica una adecuada especificación para dicho ítem en la matriz \mathbf{Q} . En resumen, un buen ítem debe tener un valor alto de π_j^* y b_j y valores bajos de r_{jk}^* .

En el RUM también se estiman dos parámetros para los examinados: α es un vector de habilidades latentes correspondiente a los estados de conocimiento de los examinados y θ_i es el parámetro de habilidad residual que mide la habilidad en atributos que no han sido especificados en la matriz \mathbf{Q} , pero que influyen en la solución correcta de los ítems. La interpretación de este parámetro es la misma que en el modelo unificado, es decir, puede implicar el uso de estrategias alternativas o la medición de habilidades de alto orden (Dibello, Stout & Roussos, 1995; Hartz, 2002).

El modelo fusión se diferencia del RUM porque incluye un parámetro (p_k) para cada atributo, dicho parámetro es un punto de corte que se utiliza para dicotomizar la variable α (que se considera continua en este modelo): si el valor de α es mayor que el punto de corte p_k , se asigna al examinado un estatus de dominio del atributo, en otro caso se establece no dominio. La estimación de parámetros para los modelos RUM y fusión se realiza mediante algoritmos MCMC que han sido implementados en un software llamado Arpeggio (Hartz, Roussos & Stout, 2002).

2.2.6. Modelo GDM

El modelo de diagnóstico general o GDM (Von Davier y Yamamoto, 2004) combina la TRI con los modelos de clases latentes. El modelo requiere la especificación de una matriz \mathbf{Q} que vincula los ítems con los atributos requeridos para resolverlos, sin embargo, el GDM permite usar matrices \mathbf{Q} no dicotómicas. Con dicho modelo se pueden analizar también ítems y atributos de naturaleza politómica, convirtiéndolo en una de las propuestas más

generales entre las que se han hecho hasta el momento. El GDM es una extensión multi-dimensional del modelo de crédito parcial (PCM; Masters, 1982). La función de respuesta del PCM es:

$$P_j(X_j = x|\theta) = \frac{\exp(x\theta + \beta_{xj})}{1 + \sum_{y=1}^m \exp(y\theta + \beta_{yj})} \quad (2.18)$$

Von Davier y Yamamoto (2004) descomponen el rasgo latente unidimensional en una combinación lineal de k atributos unidimensionales (α) para cada ítem, como se puede ver en la siguiente ecuación:

$$P_j(X_j = x|\beta_j, q_j, \gamma_j, \alpha) = \frac{\exp(\beta_{xj} + x \sum_{k=1}^K \gamma_{jk} q_{jk} \alpha_{ik})}{1 + \sum_{y=1}^m \exp(\beta_{yj} + y \sum_{k=1}^K \gamma_{jk} q_{jk} \alpha_{ik})} \quad (2.19)$$

Donde,

β_{xj} y β_{yj} son los parámetros de dificultad del ítem j .

γ_{jk} es un parámetro k -dimensional correspondiente a la pendiente de cada categoría de respuesta m .

q_{jk} es un elemento de la matriz \mathbf{Q} que especifica la correspondencia entre el ítem j y el atributo k .

α es el vector de variables latentes dicotómicas (de elementos α_{ik}) para el examinado i y el atributo A_k .

De esta manera, en la ecuación (2.19), y es un índice de las posibles puntuaciones del ítem j y m denota la máxima puntuación para ese ítem. De acuerdo con la ecuación (2.19) la probabilidad de obtener la puntuación x en el ítem j depende de los parámetros del ítem β_{xj} , β_{yj} y γ_{jk} y del perfil de habilidades del examinado α_{ik} . En este modelo, los elementos del vector α pueden asumir cualquier valor en la escala de números reales que el usuario fija en la especificación de su modelo. Von Davier (2008) implementa la estimación de parámetros para el GDM por CML en un software llamado *mdltm* utilizando el algoritmo EM. Dicho software brinda información sobre la convergencia de los parámetros, el número de iteraciones requeridas y los estadísticos de ajuste del modelo. Adicionalmente, el autor presenta una aplicación del modelo con datos empíricos de lenguaje tomados del test TOEFL (Von Davier, 2008).

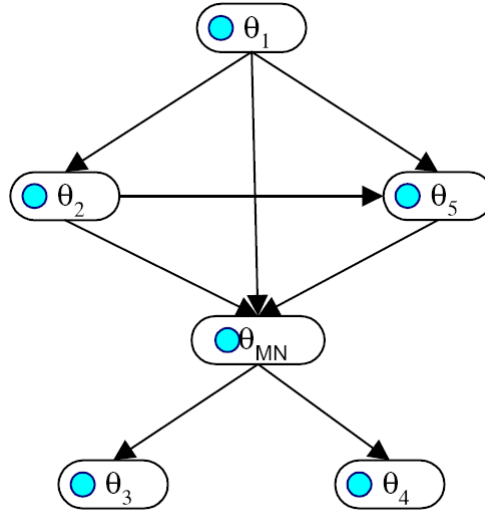
2.2.7. Redes Bayesianas

Mislevy (1995), Mislevy, Almond, Yan y Steinberg (2000) y Almond, DiBello, Moulder y Zapata-Rivera (2007) proponen el uso de las redes bayesianas para modelar la ejecución cognitiva. Una red bayesiana se representa mediante una gráfica en la que los nodos son variables y las flechas representan relaciones de influencia causal entre dichas variables. Los parámetros estocásticos de estos modelos son las probabilidades condicionadas de cada nodo dados los diferentes valores que pueden tomar sus padres, es decir, si las variables de la red son X_i , $i = 1, \dots, n$ y $PA(X_i)$ representa el conjunto de padres del nodo i , los parámetros de la red son las distribuciones condicionadas $P(X_i|PA(X_i))$, $i = 1, \dots, n$ y, en el caso de los nodos sin padres, las distribuciones a priori. Este conjunto de probabilidades es suficiente para describir totalmente la distribución de probabilidad conjunta de todas las variables de la red.

En el campo del diagnóstico cognitivo la aplicación de las redes bayesianas tiene tres fases: a) modelamiento, b) construcción de tareas y c) inferencia.

La fase de modelamiento consiste en la especificación del denominado *modelo de competencia* o *modelo del estudiante*, dicho modelo se estructura a partir de la teoría cognitiva disponible sobre el dominio de interés e incluye el conjunto de atributos latentes que caracterizan la competencia del estudiante y sus relaciones. Dicho conjunto de atributos constituyen los nodos de la red. En la Gráfica 2.1 se presenta, a modo de ilustración, la representación de un modelo de estudiante para tareas de sustracción de fraccionarios y números mixtos (Sinharay, Almond y Yan, 2004) en la que el atributo θ_1 representa la habilidad para sustraer fracciones básicas, θ_2 la habilidad de reducción o simplificación de fracciones o números mixtos, θ_3 la separación entre el número entero y la fracción, θ_4 prestar 1 del número entero en una fracción mixta y θ_5 convertir un número entero en una fracción. Un análisis previo ha mostrado que θ_3 es prerequisite de θ_4 , dicha restricción se incorpora en una variable auxiliar de 4 niveles (θ_{MN}). Un nivel de 0 en θ_{MN} indica que los examinados carecen de dominio de las dos habilidades, un valor de 1 representa a los sujetos que dominan θ_3 pero no dominan θ_4 , un valor de 2 representa a los examinados que dominan θ_4 pero no θ_3 y un valor de 3 representa a aquellos que dominan las dos habilidades. Las relaciones entre atributos pueden ser empíricas, por ejemplo, las proba-

Gráfica 2.1: Ejemplo de modelo de estudiante para tareas de sustracción de fracciones.



bilidades de dominar θ_2 dado que se domina θ_1 o lógicas, como las establecidas entre θ_3 y θ_4 . La construcción de este modelo implica varias consideraciones importantes como la decisión sobre cuáles habilidades se deben incluir, el tipo de relación entre las mismas y el nivel de detalle que se pretende modelar (habilidades gruesas o específicas).

La fase de construcción de tareas implica el diseño de situaciones en las que las diferencias en el espacio de habilidades latentes sean reflejadas en respuestas observables. Esta fase implica la construcción de un *modelo de evidencia* en el que se vinculan las variables de competencia con respuestas observables, la forma funcional de este modelo varía en cada ítem dependiendo de las filas de la matriz \mathbf{Q} , de esta manera, todos los ítems que requieren el mismo grupo de atributos conforman un solo modelo de evidencia.

La fase de inferencia implica un razonamiento desde las respuestas observadas (X) hacia las variables latentes que componen el modelo de competencia(α), que actúan como nodos padres de las variables observables que constituyen los nodos hijo. Si $p(\alpha)$ son las probabilidades a priori de los nodos padre (atributos) en la población de interés (basadas en la opinión de expertos, profesores o en investigaciones previas) y sabiendo que $p(\alpha|X)$ y $p(X|\alpha)$ son las expectativas iniciales sobre el estado de conocimiento de un estudiante, el objetivo de este procedimiento es que, una vez se cuenta con observaciones, se pueda revisar o actualizar dichas probabilidades usando una red que traza inferencias sobre α dado X usando el teorema de Bayes: $p(\alpha|X) \propto p(X|\alpha)p(\alpha)$ (Almond y Cols, 2007).

En resumen, en este enfoque se utiliza la teoría cognitiva, la opinión de jueces expertos y los resultados de estudios previos para producir distribuciones a priori de las probabilidades condicionales en un modelo multivariado de variables latentes α , correspondientes a los estados de conocimiento, y se utilizan algoritmos MCMC o redes Bayesianas para refinar dichas estimaciones a partir de las respuestas a un conjunto de ítems o tareas. Se busca también obtener perfiles de competencia de cada examinado de acuerdo al modelo de competencia estructurado en el primer paso. La inferencia por redes bayesianas es el método que se suele utilizar en los tutores inteligentes como el ACT (Corbett, Anderson, O'Brien, 1995), el LISP (Anderson, 1990) o el Quest (White y Fredericksen, 1987). Se han realizado aplicaciones de este procedimiento en modelos de solución de problemas en diversas áreas como técnicas de aviación (HYDRIVE; Gitomer, Steinberg y Mislevy, 1995), aritmética y razonamiento proporcional (Beland y Mislevy, 1992) y sustracción con números mixtos (Sinharay, Almond y Yan, 2004).

2.3. Método LSDM

Los diversos modelos presentados hasta el momento, con sus ventajas y limitaciones, proporcionan diferentes perspectivas de análisis y validación cognitiva. Por ejemplo, los modelos derivados de la TRI, como el LLTM, se enfocan en la predicción de la dificultad del ítem a partir de los atributos cognitivos, pero no profundizan en las relaciones cognitivas entre ítems y no proporcionan información diagnóstica sobre el nivel de los examinados en cada uno de los componentes. Por otra parte, los modelos de diagnóstico cognitivo proporcionan información sobre los perfiles de los examinados en el dominio de los atributos latentes usando modelos probabilísticos complejos. La mayor parte de los modelos anteriormente presentados, independientemente de sus bases matemáticas y teóricas, requieren información sobre las puntuaciones de los examinados en los ítems y no presentan evidencia independiente sobre la validez de los atributos cognitivos. Para responder a estas dos necesidades Dimitrov (2007) propone una aproximación para la validación de atributos cognitivos requeridos en la solución de ítems binarios llamado método de las distancias mínimo cuadráticas (en inglés Least Squares Distance Method-LSDM).

El método utiliza los parámetros los ítems estimados con alguno de los modelos de la

TRI (logístico de 1 parámetro, de 2 parámetros, de 3 parámetros o el modelo de Rasch) y la matriz de especificación cognitiva \mathbf{Q} para estimar la probabilidad de dominio de los atributos (A_k) en niveles de habilidad previamente establecidos; es por este motivo que para usar el método no se requiere información sobre las puntuaciones de los examinados, siempre y cuando se cuente con estimaciones de los parámetros de algún modelo de la TRI. Es por este motivo que el método se puede usar para la validación de atributos en bancos de ítems de los que se tenga información previa y para analizar atributos en estudios que reporten las estimaciones de los parámetros de los ítems.

Al igual que en otros modelos ya vistos, en el LSDM $A_1, A_2 \dots A_k$ denotan las operaciones o atributos cognitivos asociados a un test de n ítems binarios y \mathbf{Q} es la matriz de especificación cognitiva con elementos $q_{jk} = 1$ si la respuesta correcta del ítem j requiere el dominio del atributo A_k y $q_{jk} = 0$ en otro caso. El LSDM asume una relación conjuntiva entre los atributos según la cual la probabilidad de resolver correctamente un ítem j para una persona i con habilidad θ_i , denotada por P_{ij} equivale al producto de las probabilidades de dominio de los atributos requeridos por dicho ítem:

$$P_{ij} = \prod_{k=1}^K [P(A_k = 1|\theta_i)]^{q_{jk}} \quad (2.20)$$

Donde,

P_{ij} es la probabilidad de responder correctamente el ítem j en el nivel de habilidad θ_i .

$P(A_k = 1|\theta_i)$ es la probabilidad de dominio del atributo A_k en el nivel de habilidad θ_i .

q_{jk} es el elemento de la matriz \mathbf{Q} para el ítem j y el atributo A_k .

A diferencia de los modelos componenciales presentados anteriormente, en el LSDM no se estiman las probabilidades de dominio de los atributos con un modelo de TRI o un modelo de clase latente, ni se actualizan por medio de redes bayesianas, sino que se estiman directamente utilizando un método proveniente del álgebra lineal en el que se busca la solución a un sistema de ecuaciones conformado por la forma matricial de los elementos de la ecuación 2.20. Los pasos para la aplicación de esta metodología son:

1. Se fija un número razonable de niveles de habilidad en la escala logística.
2. Se obtienen los parámetros de la TRI: se pueden obtener mediante reportes de estudios previos o mediante calibración, si se dispone de datos. En el último caso es

necesario verificar el ajuste del modelo que se está empleando.

3. Teniendo los parámetros de los ítems y los niveles fijos de habilidad, en la escala logística, se encuentra la probabilidad (P_{ij}) de resolver correctamente el ítem j en cada nivel de habilidad previamente fijado aplicando el modelo TRI correspondiente.
4. Se toma el logaritmo neperiano a ambos lados de la ecuación (2.20) produciendo el siguiente sistema de ecuaciones lineales:

$$\mathbf{L} = \mathbf{Q} \cdot \mathbf{X} \quad (2.21)$$

Donde,

\mathbf{L} es un vector ($n \times 1$) con elementos $\ln(P_{ij})$ (conocido).

\mathbf{Q} es la matriz ($n \times k$) de especificación cognitiva (conocida).

\mathbf{X} es el vector ($k \times 1$) de incógnitas con elementos $\ln[P(A_k = 1|\theta_i)]$.

Dado que una de las características de este tipo de aplicaciones es que suele haber más ítems que atributos, el sistema de ecuaciones lineales resulta sobredeterminado teniendo más ecuaciones que incógnitas ($\mathbf{L} \cong \mathbf{Q} \cdot \mathbf{X}$), por esta razón, para encontrar una solución, el autor propone usar el método de la distancia mínimo-cuadrática por medio del cual se encuentra el vector \mathbf{X} que minimiza la norma Euclídea de $\mathbf{Q} \cdot \mathbf{X} - \mathbf{L}$. Como la matriz \mathbf{Q} es siempre de rango completo (no puede tener un vector columna de 0 porque cada atributo es requerido por lo menos en un ítem), existen dos posibles vías para minimizar la norma; una de ellas es la descomposición ortogonal de la matriz \mathbf{Q} , y la otra es la solución clásica usando la pseudoinversa: $(\mathbf{Q}^T \mathbf{Q})^{-1} \mathbf{Q}^T \mathbf{L}$, las dos aproximaciones conducen a los mismos resultados (Lawson y Hanson, 1974).

Una vez se ha obtenido el vector solución \mathbf{X} , se encuentra la probabilidad de dominio del atributo A_k en cada uno de los niveles fijos de habilidad θ_i que corresponde al exponencial del vector solución: $\exp(\mathbf{X}_k) = P(A_k = 1|\theta_i)$. La representación gráfica de este vector corresponde a la Curva de Probabilidad del Atributo (CPA) para el atributo A_k . Un asunto técnico a resolver en este punto es que los elementos del vector solución \mathbf{X} deben ser positivos ya que P_{ij} debe estar siempre entre 0 y 1, para ello se incluye una restricción de negatividad en la solución del vector que no afecta los resultados de la estimación (Dimitrov, 2007).

El último paso del procedimiento es emplear de nuevo la ecuación (2.20) para obtener las CCIs recuperadas con el LSDM, mediante la multiplicación de las probabilidades de dominio de los atributos $[P(A_k = 1|\theta_i)]$ estimadas por el método. La comparación entre las CCIs obtenidas originalmente con la TRI y las recuperadas con el LSDM proporcionan evidencia sobre la calidad de los vínculos entre ítem y atributos que han sido especificados en \mathbf{Q} . Aunque suelen presentar la misma forma es importante diferenciar la CCI de la CPA pues la primera representa la probabilidad de respuesta correcta a un ítem y la segunda la probabilidad de dominio de un atributo A_k en niveles fijos de habilidad.

Dimitrov (2007) propone tres criterios para acopiar evidencia sobre la validez de la estructura cognitiva representada en \mathbf{Q} :

1. Residuales pequeños: Los valores LSD (Least Squares Distance) son los residuales obtenidos tras la minimización del vector $\|\mathbf{Q} \cdot \mathbf{X} - \mathbf{L}\|$ y deben ser pequeños para indicar una solución apropiada para el vector \mathbf{X} , cuanto más pequeños son estos valores más precisa es la explicación de la ejecución en el ítem por la conjunción de los atributos propuestos en dicho nivel de habilidad.
2. Comportamiento de las CPAs: si la especificación de \mathbf{Q} es correcta, las CPAs deben exhibir la propiedad de monotonidad pues cuanto mayor sea la habilidad latente de un examinado (parámetro de habilidad obtenido con la TRI) es más probable que domine los atributos componentes de la tarea. Las CPAs también proporcionan información sobre la discriminación y dificultad relativa de los atributos. Esta información puede ser útil para evaluar la validez de la estructura cognitiva: si la diferencia en la dificultad de los atributos es suficientemente grande, las CPAs no se cruzarán y los atributos pueden ser ordenados según su dificultad para luego ser comparados con las expectativas que se tenían de ellos, procedente de la teoría cognitiva, la opinión de expertos o las probabilidades de atributos definidas a priori. Por ejemplo, en el caso de operaciones aritméticas con números enteros, se espera que los atributos de suma y resta con números naturales sean más fáciles que los atributos que involucren operaciones con números enteros, estas expectativas se pueden contrastar con los resultados del LSDM y evaluar el comportamiento lógico de las CPAs pues grandes diferencias entre las expectativas del modelo cognitivo y los resultados

del LSDM pueden significar errores de especificación en la matriz \mathbf{Q} . Por medio de las CPAs se puede realizar también una validación cruzada de los resultados del LSDM con otros modelos, como el LLTM o el DINA, mediante la comparación del ordenamiento de los atributos según su dificultad: en el caso del LLTM se pueden ordenar según los parámetros básicos y luego compararlos con el orden obtenido en el LSDM, en el caso del DINA se puede hacer lo mismo usando los parámetros de prevalencia de los atributos.

3. Recuperación de las CCIs: la comparación de las CCIs obtenidas con la TRI y las recuperadas mediante el método LSDM proporciona evidencia sobre la validez de la estructura cognitiva pues una buena recuperación indica que la ejecución en el ítem está bien explicada por los atributos propuestos. Dimitrov (2007) ha cuantificado este criterio por medio de la diferencia media absoluta (en inglés, Mean Absolute Difference o MAD). Se puede calcular un valor MAD para cada ítem j pues corresponde a la media (a través de los m niveles fijos de habilidad) de las diferencias absolutas entre la CCI obtenida por la TRI y la estimada por el LSDM, como se puede observar en la siguiente ecuación:

$$MAD_j = \frac{\sum_{m=1}^M |P_{jm} - \hat{P}_{jm}|}{M} \quad (2.22)$$

Donde,

M es el número de niveles de habilidad (θ_i) previamente fijados.

P_{jm} es la probabilidad para el ítem j en el nivel de habilidad m obtenida con la TRI.

\hat{P}_{jm} es la probabilidad para el ítem j en el nivel de habilidad m , estimada por el LSDM.

Idealmente un valor $MAD = 0$ indicaría una perfecta recuperación. El autor presenta la siguiente regla empírica de clasificación para los valores de MAD basada en algunas aplicaciones del LSDM con datos reales: muy buena recuperación ($0 \leq MAD \leq 0.02$), buena ($0.02 < MAD \leq 0.05$), algo buena ($0.05 < MAD \leq 0.1$), algo pobre ($0.1 < MAD \leq 0.15$), pobre ($0.15 < MAD \leq 0.2$) y muy pobre ($MAD > 0.2$).

Dimitrov (2007) ha realizado dos aplicaciones del LSDM: en el primer estudio utiliza los parámetros obtenidos con el modelo de Rasch para un test de Matemáticas con 19 ítems y propone una matriz \mathbf{Q} compuesta por 8 atributos para explicar la ejecución de los sujetos en dichos ítems. Tanto la matriz \mathbf{Q} como los parámetros de los ítems habían sido reportados

en un estudio de Fisher (1973). En esta primera aplicación, el autor encontró valores LSD entre 0.007 y 0.082. Dichos valores son mas altos en los niveles de habilidad bajos y van decreciendo a medida que aumenta la habilidad. Las CPAs presentan el comportamiento monótono esperado y corresponden con las expectativas provenientes de la teoría cognitiva. Dimitrov aplica también a sus datos el LLTM y realiza una validación cruzada de los resultados de los dos modelos, evidenciando que el orden de la dificultad establecido por el LSDM se empareja perfectamente con el orden de los parámetros básicos estimados con el LLTM. Por otra parte, las CPAs exhiben un correcto comportamiento en su discriminación pues los atributos fáciles (A_4) discriminan mejor a los examinados de baja habilidad, mientras que los más difíciles, como A_2 y A_7 , discriminan mejor a los de habilidad alta. En cuanto a la recuperación de las CCIs el autor encontró que sólo cuatro ítems presentan mala recuperación de acuerdo a la regla empírica anteriormente mencionada.

La segunda aplicación se ha centrado en un test de lectura con cinco atributos. Dado que el autor no disponía de parámetros de los ítems, el primer paso ha sido estimarlos encontrando que el modelo de TRI con mejor ajuste es el logístico de dos parámetros. Los valores residuales (LSD) fueron más altos que en la situación anterior (entre 0.05 y 0.208) reflejando una menor precisión. Como era de esperarse, la recuperación de las CCIs ha sido también peor en este caso (dos de los 10 ítems muestran una mala recuperación). En este ejemplo, el autor muestra otra aplicación práctica del modelo: la comparación entre grupos de examinados en cuanto a las probabilidades de dominio de los atributos, en particular, presenta los resultados de la comparación de la ejecución de las tareas entre hombres y mujeres.

Además de los estudios reportados por Dimitrov (2007) también se han realizado otras dos aplicaciones del LSDM. Ma, Çetin y Green (2009) realizan una aplicación del LSDM en la sub-prueba de Matemáticas de la evaluación internacional TIMSS. Para ello han utilizado una muestra de 30826 estudiantes de octavo grado de siete regiones de Turquía. Los autores han analizado 25 ítems de elección múltiple que incluyen integrales, fracciones, álgebra, geometría, probabilidad y estadística básica. Se asume una estructura cognitiva de 20 atributos cognitivos subyacentes a la ejecución de los ítems del test y dicha estructura se ha dividido en tres áreas: 6 atributos de contenido/conocimiento, 7 procesos cognitivos y 7 habilidades. El LSDM se aplica a cada grupo de atributos por separado

encontrando que: a) en los atributos de contenido el orden de las dificultades relativas y la forma de las CPAs corresponden a lo esperado y sólo uno de ellos (*conceptos y operaciones con fracciones y decimales*) se encuentra un comportamiento extraño en el que las probabilidades de dominio son una constante de 1 en los niveles de habilidad superiores a -2; b) en cuanto a los atributos de procesos cognitivos los autores encuentran dos CPAs con comportamientos inesperados, la curva correspondiente al atributo *manejo de datos y procedimientos* no es monótona creciente y las probabilidades de dominio del atributo *lectura cuantitativa y lógica* son 1 para todos los niveles de rasgo; c) en los atributos de habilidades encontraron nuevamente dos atributos con comportamiento inesperado: la curva correspondiente al atributo *comparación de dos o más entidades* no es monótona y las probabilidades de dominio del atributo *entendimiento de cuestiones verbales* son 1 para todos los niveles de habilidad. Por dicho motivo los autores decidieron reespecificar el modelo cognitivo eliminando los atributos con malas propiedades, tras ejecutar nuevamente el método, encontraron que la estimación de probabilidades de los atributos restantes permaneció igual.

En cuanto a la recuperación de las CCIIs los autores reportan que el 80 % de los ítems fueron recuperados correctamente con un criterio de $MAD < 0.1$. Los valores LSD máximos para cada condición han sido 0.36 (atributos de habilidades), 0.34 (atributos de procesos cognitivos) y 0.26 (habilidades de contenido) indicando que los atributos de contenido explican mejor los ítems que los restantes atributos (Ma, Çetin y Green, 2009). Finalmente los autores realizan comparaciones de la ejecución en los atributos entre grupos de hombres y mujeres y en las diversas regiones de Turquía encontrando que mujeres y hombres presentan similar desempeño en los atributos y que los estudiantes de las regiones del este de Turquía presentan un menor rendimiento en todos los atributos.

Romero, Ordóñez, López y Navarro (2009) analizan la estructura cognitiva de la competencia científica en PISA-2006 por medio del LSDM. Para ello utilizan una muestra de 1538 alumnos de España que respondieron el cuadernillo uno de PISA-2006. Los autores han analizado 18 ítems y 3 atributos cognitivos propuestos por la OCDE para explicar la ejecución de los ítems de ciencias. Los resultados muestran que las CPAs exhiben el comportamiento lógico esperado: a) son monótonas crecientes, indicando que a medida que aumenta la habilidad aumenta también la probabilidad de dominio de los atributos,

y b) el orden de la dificultad de los atributos coincide con lo esperado de acuerdo a la teoría. El proceso más fácil ha resultado ser A_3 : utilizar pruebas científicas, seguido por A_2 : explicar fenómenos científicamente y el más difícil ha sido A_1 : identificar cuestiones científicas. En cuanto a la recuperación de las CCIs, los autores reportan que el 75 % de los ítems se encuentran bien recuperados con un criterio de $MAD < 0.1$. En este estudio también se analiza si existen diferencias significativas en el desempeño de los atributos entre comunidades autónomas españolas encontrando diferencias entre el País Vasco y las comunidades Aragón y Navarra en todos los atributos.

2.4. Otros métodos de validación

2.4.1. Procedimientos empleados en la Psicología Cognitiva

En el marco de la Psicología Cognitiva se han propuesto diversos métodos para evaluar la calidad de los vínculos entre las tareas y los componentes hipotéticos o atributos que subyacen a su resolución. La mayoría de estas aproximaciones son de carácter experimental, por ejemplo, mediante el estudio de tiempos de respuesta o latencias (Sternberg y Weil, 1980) y protocolos o reportes verbales (Lohman, 1989; Leighton y Gierl, 2007). Este tipo de estudios aportan evidencia fundamental en el proceso de explorar y encontrar atributos relevantes para el desarrollo de las tareas e ítems, sin embargo son poco informativos en la fase de revisión de la estructura cognitiva tras la aplicación de los modelos componenciales.

2.4.2. Procedimientos basados en modelos componenciales

La mayor parte de los modelos componenciales basados en la TRI incluyen concepciones sobre el proceso de validación de los atributos cognitivos y procedimientos para acopiar evidencia de la misma.

Una de las líneas de procedimientos es la usada por Embretson (1984), quien realiza diversas verificaciones de la estructura cognitiva en la que apoya sus estudios: en primer lugar, evalúa el ajuste mediante pruebas de RV de los modelos componenciales que utiliza (LLTM - Whitely y Schneider, 1981; MLTM – Whitely, 1980 y GLMT – Embretson, 1984), luego compara el ajuste entre los diversos modelos y finalmente compara la bondad de ajuste del modelo cognitivo original con diversas alternativas (agregando o eliminando

atributos con criterios teóricos). Esta autora también realiza regresión por etapas con el objetivo de evaluar el decrecimiento de los valores R^2 a medida que va excluyendo componentes de la ecuación (atributos cognitivos). Un decrecimiento pequeño en el R^2 implica que el atributo no es demasiado importante para resolver la tarea. Embretson también pone a prueba la estructura cognitiva por medio de la comparación de parámetros (tanto de los ítems como de los examinados) obtenidos en las respuestas a los ítems totales con los parámetros obtenidos en las sub-tareas. En particular, para el caso de los examinados, la autora presenta las correlaciones entre la puntuación total y las habilidades estimadas en cada componente y, para el caso de los ítems estudia las correlaciones entre los parámetros de la tarea compuesta y de cada subtarea.

Al igual que Embretson, Maris (1999) propone probar la que denomina *hipótesis cognitiva* por medio de la comparación entre la predicción de las respuestas de los examinados usando el modelo componencial y las respuestas empíricas. La diferencia entre las dos aproximaciones es que Embretson utiliza un procedimiento en dos pasos, en el cual primero estima los parámetros de las subtarear y luego utiliza las estimaciones para predecir las respuestas a los ítems. Maris une estos dos pasos en una sola estimación en la cual no es necesario tener las respuestas empíricas de las subtarear. Los procedimientos empleados por Maris y Embretson (evaluar el ajuste de los modelos y estudiar correlaciones entre los parámetros de las tareas compuestas y las subtarear) aportan importantes piezas de evidencia sobre la validez de la estructura cognitiva, sin embargo, estos procedimientos presentan algunos inconvenientes: a) en algunos modelos componenciales no es posible aplicar la prueba de RV excepto para medidas con pocos ítems y muchos examinados (Sinharay, Almond, Yan, 2004), b) no permiten conocer si la falta de ajuste es causada por una especificación errónea de la matriz \mathbf{Q} o por otras fuentes de error como la estimación de los parámetros o una mala elección del modelo y c) mediante la evaluación del ajuste solo se obtiene información global sobre la calidad de la matriz \mathbf{Q} pero se desconoce cuáles son los vínculos particulares ítem-atributo que están incorrectamente especificados.

En el contexto de los modelos de diagnóstico cognitivo también se han propuesto algunas ideas y procedimientos para acopiar evidencia sobre la validez de los atributos cognitivos. En el marco del RSM (Tatsuoka, 1995) una alta tasa de examinados sin clasificar en algún estado de conocimiento es un indicador de mala especificación de \mathbf{Q} . Una

matriz \mathbf{Q} correctamente especificada deberá presentar tasas de clasificación sobre el 90 %.

De la misma manera, Rupp y Templin (2008) estudian los efectos de la introducción de errores de especificación en la matriz \mathbf{Q} sobre la recuperación de parámetros y las tasas de clasificación cuando se usa el modelo DINA. Los autores encuentran una sobreestimación de los parámetros de desliz (s) cuando se eliminan atributos de la matriz \mathbf{Q} verdadera y subestimación de los parámetros de azar (g) cuando se adicionan atributos irrelevantes. Los autores también encuentran tasas altas de clasificación errónea en las clases latentes que contienen combinaciones de atributos en los que se han introducido errores de especificación. Estos resultados muestran los efectos de algunas fuentes de invalidez sobre la estimación de los parámetros del modelo, sin embargo, dichos resultados solo pueden aplicarse al modelo DINA, siendo poco generalizables.

En el modelo unificado (Dibello, Stout y Roussos, 1995) se tienen en cuenta las habilidades no especificadas en \mathbf{Q} mediante la introducción de un parámetro de habilidad unidimensional; mediante dicho parámetro se puede conocer si \mathbf{Q} contiene todos los componentes o atributos relevantes para resolver la tarea. En este modelo se define también un parámetro de ítem que se puede usar para diagnosticar si los ítems están bien representados por los atributos especificados en la matriz \mathbf{Q} (Hartz, 2002). En el modelo unificado también se incluyen otras dos fuentes de error: el uso de estrategias alternativas y la *positividad*, que refleja el grado en el cual un examinado que domina un atributo realmente lo usa para resolver la tarea. Como se ha visto, el modelo unificado permite acopiar información sobre varios niveles del proceso de validación: permite probar si los componentes especificados en \mathbf{Q} son suficientes para explicar el desempeño de los examinados y también evaluar las relaciones ítem-atributo; adicionalmente se tiene en cuenta también el uso de estrategias alternativas. A pesar de que es un modelo que brinda gran cantidad de información tiene el inconveniente de que sus parámetros no son estimables pues el modelo no está identificado.

Finalmente, en el enfoque por redes bayesianas (Mislevy, 1995) también hay aproximaciones a la validez. Desde dicha perspectiva es muy importante el diseño del proceso de medición que incluye una cuidada especificación de las probabilidades de dominio previas. La recuperación de dichas probabilidades y los intervalos de confianza establecidos para las mismas aportan evidencia sobre la validez de la estructura cognitiva, pues cuanto más

se acerque la probabilidad actualizada a la definida a priori, mayor evidencia de que la estructura considerada es válida.

En resumen, la mayoría de los modelos componenciales que han sido analizados incluyen propuestas sobre el proceso de validación de la estructura cognitiva y métodos para acopiar información sobre la exactitud de la matriz \mathbf{Q} ; los enfoques son diversos y contribuyen al proceso integral de validación aportando evidencia complementaria. Sin embargo la mayor parte de estas ideas son dependientes del modelo aplicado y son poco generalizables a otros modelos u otras aplicaciones.

2.4.3. Procedimientos de validación independientes del modelo componencial

2.4.3.1. Quadratic Assignment (QA)

En 1993 Medina-Díaz propone la técnica QA para el análisis de las estructuras cognitivas teóricas. El QA es un procedimiento estadístico que evalúa las relaciones entre la matriz de proximidad (\mathbf{P}) y la matriz estructural (\mathbf{C}). La matriz \mathbf{P} contiene alguna medida empírica de proximidad entre parejas de ítems basada en las respuestas de los examinados. En el caso del modelamiento cognitivo, \mathbf{P} indica el grado en el cual una pareja de ítems han sido resueltos usando la misma combinación de atributos. Por su parte, la matriz \mathbf{C} corresponde a la matriz de incidencia que contiene el número de atributos que una pareja de ítems tienen en común según la estructura cognitiva hipotética del test. El análisis QA confirmatorio prueba si la estructura especificada en \mathbf{C} está representada en \mathbf{P} . Se utiliza un índice Γ para medir la fuerza de la asociación entre \mathbf{C} y \mathbf{P} :

$$\Gamma = \sum_x^N \sum_y^N [p(x, y)c(x, y)] \quad (2.23)$$

Donde,

x e y son dos ítems diferentes.

$p(x, y)$ y $c(x, y)$ son elementos de \mathbf{P} y \mathbf{C} respectivamente para el par de ítems.

La hipótesis nula es que las filas y columnas de \mathbf{C} se distribuyen al azar con respecto a la estructura especificada en \mathbf{P} . Para probar esta hipótesis se compara el valor de Γ con un valor crítico derivado de una distribución de referencia que es obtenida aleatoriamente.

Si el valor de Γ cae en la región crítica ($p > 0.05$), se rechaza la hipótesis nula indicando que \mathbf{P} refleja la estructura cognitiva representada en \mathbf{C} .

Medina-Díaz (1993) presenta una aplicación empírica de este procedimiento analizando con el LLTM los datos de un test de álgebra con 29 ítems y 8 atributos hipotéticos. La técnica QA se ha empleado con el objetivo de determinar si la estructura cognitiva hipotética estaba presente en las respuestas de los examinados. Para determinar la matriz \mathbf{P} , la autora examinó cada paso llevado a cabo por los estudiantes para resolver los ítems, ésta constituye la principal desventaja del procedimiento QA pues requiere información empírica del desempeño de los sujetos en cada atributo. En primer lugar, se genera una variable según la cual se asigna un valor de 1 si un examinado usa un atributo y un valor de 0 si no lo usa, partiendo de dicha asignación se calcula la matriz \mathbf{P} de acuerdo a la siguiente expresión (Medina-Díaz, 1993):

$$P(x, y) = \sum_{i=1}^I \sum_{k=1}^K (d_{kx} d_{ky}) \alpha_k \quad (2.24)$$

Donde,

d_{kx} y d_{ky} son las variables binarias (1-0) asociadas con cada atributo.

α_k es el parámetro básico estimado con el LLTM para el atributo A_k .

K es el número de atributos.

I es el número de examinados.

Así, en el ejemplo de Medina-Díaz, \mathbf{P} es una matriz simétrica de 29×29 , sus elementos resumen los 235 patrones de respuesta a cada par de ítems en función de los 8 atributos. La matriz estructural (\mathbf{C}) también es una matriz simétrica de 29×29 y sus elementos contienen la suma de los productos de los vectores de la matriz que contiene la estructura cognitiva (matriz \mathbf{W} en el modelo LLTM):

$$c(x, y) = \sum_{k=1}^K w_{kx} w_{ky} \quad (2.25)$$

En la ecuación (2.25) se puede ver que los elementos de \mathbf{C} reflejan las relaciones hipotéticas entre ítems basadas en la matriz \mathbf{W} pues corresponden al número común de atributos requeridos para resolver cada par de ítems, por lo cual, son medidas de similitud entre la pareja de ítems. Los elementos diagonales de \mathbf{P} y \mathbf{C} se fijan a 0.

Los resultados encontrados por Medina-Díaz en el test de álgebra evidencian que existe un acuerdo significativo entre **P** y **C**, por lo tanto, los patrones de respuesta reflejan la estructura de la matriz **C**. En una aplicación posterior de la metodología QA, López y Elosua (2002) encontraron también un acuerdo significativo entre las matrices **C** y **P** $p < 0.001$ en un test de 21 ítems de adición y sustracción de fracciones y 6 atributos cognitivos. Además del estudio mencionado no se han encontrado otras aplicaciones de QA en este contexto.

2.4.3.2. Análisis Clúster

Corter (1995) propone que los modelos discretos de similaridad de ítems, como el análisis de conglomerados y técnicas de árbol, pueden ser herramientas más apropiadas para acopiar evidencia sobre la validez de estructuras cognitivas que los modelos geométricos tradicionalmente empleados en la Psicometría como el análisis factorial o el análisis en componentes principales. El autor propone que dichas técnicas pueden ser útiles tanto de manera exploratoria, cuando no se conoce una estructura cognitiva teórica, o confirmatoria, para dar significado adicional o validar la caracterización de un test en términos de atributos cognitivos específicos. El autor se basa en los modelos básicos de similaridad, de naturaleza discreta, en los que las correlaciones entre ítems pueden considerarse como una medida de proximidad; la correlación será alta cuando dos ítems requieren el mismo tipo de conocimiento es decir los mismos atributos cognitivos. Corter enfoca su trabajo en las metodologías de conglomerados tanto jerárquicas (de árbol) como no jerárquicas. Este tipo de métodos tienen algunas ventajas sobre otras técnicas al representar las correlaciones entre ítems en test de diagnóstico cognitivo: en primer lugar ofrecen la posibilidad de caracterizar los atributos mediante parámetros (pesos de los clúster), en segundo lugar han mostrado un mejor ajuste para este tipo de medición y finalmente permiten identificar cuáles son los atributos que comparte un grupo de ítems (comunalidad) y cuales son propios de un ítem (singularidad).

Modelos jerárquicos.

Tradicionalmente se han empleado dos modelos jerárquicos o de árbol: el ultramétrico y el aditivo, el primero corresponde al comúnmente denominado *dendograma*, que a dife-

rencia de los árboles aditivos tienen la restricción de que los nodos deben ser equidistantes de la raíz. La ventaja del modelo aditivo para la validación de estructuras cognitivas es que tiene la capacidad de representar la singularidad de los ítems, es decir, el grado en el cual los ítems tienen una única característica no compartida con otros ítems en el test. Los árboles ultramétricos pueden ser interpretados en términos de las características comunes de un grupo de ítems, mientras que los aditivos son interpretados en términos de características distintivas. En un árbol ultramétrico rotado cada brazo determina un grupo particular de ítems y su longitud indica la importancia de la característica que es común a los ítems de ese grupo, mientras en el aditivo representa el peso total de las características distintivas de los ítems de dicho grupo en relación a los demás. En el modelo aditivo el lugar de la raíz es arbitrario. Cuando se aplican estas técnicas a la matriz de correlaciones de ítems de un test diagnóstico estas características comunes pueden ser identificadas con operaciones cognitivas requeridas para la solución correcta de los ítems. Por ejemplo, la existencia de un grupo conformado por tres ítems evidencia que éstos tienen uno o más componentes cognitivos comunes, la importancia de dichos atributos en la ejecución de la tarea es indicada por la longitud del arco correspondiente al clúster dado. Corter (1995) afirma que los árboles aditivos son más útiles para analizar las correlaciones entre los ítems por su capacidad para representar su singularidad, sin embargo, la interpretación de un árbol aditivo rotado es menos sincera que la de un árbol ultramétrico porque un arco en el árbol aditivo puede representar tanto características del grupo de ítems de la rama como características compartidas con ítems que no están en la rama. Un ítem con poca singularidad estará cerca de la raíz del árbol y un ítem único estará alejado de la raíz.

Modelos no jerárquicos.

Estos modelos también representan la similaridad de dos ítems en términos del clúster al cual pertenecen y no restringen el número de clústeres. Existen dos modelos de este tipo que han sido usados comúnmente en Psicología: el modelo clúster aditivo (ADCLUS; Arabie y Carroll, 1980) y el modelo de árbol extendido (EXTREE; Corter y Tversky, 1986). El primero representa la similaridad de dos ítems como la suma de pesos de los conglomerados a los que pertenecen ambos ítems. Si se asume que cada clúster representa un componente cognitivo entonces ADCLUS es un modelo de características comunes y

la similaridad entre la pareja de ítems corresponde a la suma simple de los pesos de cada componente que ambos comparten. El programa más usado para ajustar estos modelos es el MAPCLUS (Arabie y Carroll, 1980), este programa genera una serie de clústeres formados por un grupo de ítems y un peso que indica la importancia del clúster, el usuario puede introducir el número de clústeres que quiere encontrar. Con ADCLUS se puede reconstruir la matriz \mathbf{Q} pues cada clúster consiste en un grupo de ítems que incorporan un atributo y puede ser representado en un vector binario en el que el valor 1 indica que el ítem requiere el atributo y 0 en caso contrario. ADCLUS al ser un modelo de características comunes no representa la singularidad del ítem. El modelo que sí lo hace es el EXTREE, este modelo añade segmentos marcados que representan clústeres que cortan la estructura básica del árbol (y no se anidan con otros clústeres). Por ejemplo el atributo A_1 puede presentarse en dos arcos: uno que incluye el ítem 3 y otro que tiene los ítems 7 y 8, esto quiere decir que los tres ítems mencionados tienen en común el atributo A_1 . La distancia entre dos ítems es calculada como la longitud de la ruta entre ellos, excepto si dos ítems tienen en común un segmento de marcación, pues estos no entran en las distancias de longitud entre los dos ítems. El árbol extendido presenta muchas de las ventajas gráficas de la representación aditiva por ejemplo la importancia de los clúster se indica por la longitud de los arcos del árbol. También la singularidad del ítem puede ser modelada por la longitud de los arcos, pero el EXTREE tiene la ventaja de que representa grupos no anidados de atributos cognitivos, característica muy útil para los casos en que los atributos son independientes. Finalmente, Corter (1995) afirma que EXTREE puede ser el método más apropiado para analizar la estructura correlacional de los resultados de test diagnósticos porque: a) es un modelo discreto que representa las correlaciones inter-ítem en términos de un grupo de atributos discretos, b) tiene la capacidad de representar grupos de atributos cognitivos no anidados y c) permite estudiar la singularidad del ítem en términos de sus características específicas.

En resumen, Corter (1995) propone aplicar el análisis clúster a una matriz de correlaciones entre los ítems de una medición componencial con el objetivo de encontrar un grupo de clústeres que, presumiblemente, correspondan a los atributos cognitivos hipotetizados y de esta forma comparar la matriz \mathbf{Q} hipotética con la matriz \mathbf{Q} reconstruida (aproximación confirmatoria) o examinar la forma en que se pueden agregar los ítems de acuerdo

a sus características comunes y singulares (aproximación exploratoria). Este método tiene la desventaja de que algunos clústeres pueden corresponder a otros tipos de características comunes de los ítems como contenidos comunes o posición en el test. Además, descontando el ejemplo presentado por el autor, no se han realizado aplicaciones de esta metodología ni se ha probado su eficiencia y propiedades en estudios de simulación. Otro inconveniente de la presente metodología es que deja por fuera consideraciones sobre la naturaleza, relación y suficiencia de los atributos.

2.4.3.3. Ecuaciones estructurales (SEM)

Dimitrov y Raykov (2003) proponen el uso de modelos de ecuaciones estructurales (SEM) para el estudio de la validez de estructuras cognitivas en el contexto de mediciones psicométrico-cognitivas. El análisis SEM evalúa si las relaciones observadas entre los ítems son compatibles con las relaciones expresadas en el modelo psicológico propuesto para explicar cómo se resuelven. El procedimiento de validación mediante SEM propuesto por estos autores sigue los siguientes pasos:

1. Expresar la matriz \mathbf{Q} en términos de una matriz de subordinaciones entre ítems \mathbf{S} . Se dice que un ítem I_j está subordinado a un ítem I_i si y sólo si las operaciones cognitivas requeridas para su correcta resolución son un subconjunto propio de las operaciones requeridas por el ítem I_i . En la matriz \mathbf{S} , $s_{ij} = 1$ indica subordinación del ítem I_j al I_i ; en otro caso, $s_{ij} = 0$.
2. Representar gráficamente las subordinaciones definidas en la matriz \mathbf{S} . Cada ítem se representa en un diagrama de rutas (*path*) en el que las flechas I_j al I_i indican subordinación del ítem I_j al I_i .
3. Evaluar el ajuste del modelo mediante los procedimientos de la metodología SEM.

Para I_i se asume que existe una variable latente continua (η_i) que representa la habilidad requerida en su correcta solución. Existe también un valor del umbral de habilidad (τ_i) cuya superación implica el acierto del ítem, de forma que $X_i = 1$, si $\eta_i \geq \tau_i$ y $X_i = 0$, si $\eta_i < \tau_i$.

Partiendo de lo anterior, la ecuación estructural del modelo es:

$$\eta = B\eta + \zeta \quad (2.26)$$

donde,

η es el vector de variables latentes que representan las habilidades para resolver los ítems.

B es la matriz de relaciones entre las habilidades.

ζ es el vector de residuales.

Los valores de B representan el grado en que la habilidad para acertar un ítem está relacionada linealmente con la de otro(s) ítem(s) subordinados a él. Por tanto, el enfoque SEM proporciona una validación general de la estructura cognitiva y también de las relaciones de subordinación entre ítems, indicando cuáles rutas se pueden excluir o añadir.

Además de aplicar la metodología SEM, los autores proponen la triangulación, un procedimiento adicional que consiste en comprobar que la dificultad de los ítems aumenta a medida que se pasa de un ítem I_i a otro I_j (si I_i está subordinado a I_j) pues al incrementarse el número de procesos involucrados, la dificultad del ítem deberá ser mayor.

Dimitrov y Raykov (2003) presentan un ejemplo de aplicación de esta metodología en un test de 15 ítems de álgebra y 10 atributos cognitivos representados en la matriz \mathbf{Q} . Los autores reportaron un ajuste aceptable del modelo: $\chi^2 = 162.81$ con 84 grados de libertad, RMSEA = 0.058 con un intervalo de confianza al 90 %: [0.045-0.071]. En este modelo encontraron una ruta no significativa (la que relaciona el ítem 14 con el 15), dicha ruta fue eliminada y nuevamente se valoró el ajuste del modelo encontrando que no hubo un decrecimiento significativo en el estadístico chi cuadrado ($\chi^2 = 163.89$, con 85 grados de libertad). Usando este modelo como solución final los autores siguieron encontrando un par de relaciones no significativas. Los resultados de la triangulación en este estudio confirman los encontrados mediante el modelo SEM.

En otra aplicación de esta metodología (Romero, Ponsoda y Ximenez, 2006), hacen un estudio de validación de la estructura cognitiva de un test de operaciones aritméticas con números enteros. Los resultados muestran relaciones de subordinación fuertes y positivas entre los ítems en algunas de las rutas, apoyando parcialmente la estructura cognitiva de seis atributos propuesta por los autores, los resultados del proceso de triangulación son

consistentes con las relaciones de subordinación expresadas en el modelo gráfico.

Una ventaja de la utilización del enfoque SEM con respecto a otras técnicas como QA, es que no requiere información empírica sobre si el examinado ha aplicado o no cada paso en la solución del ítem, basta con saber si lo ha resuelto bien o mal, lo que constituye un ahorro en tiempo y costes. Otra ventaja es que no está ligado a los supuestos específicos del modelo componencial empleado o de los modelos de la TRI. No obstante, esta técnica también presenta limitaciones. La principal es que requiere el cumplimiento de los restrictivos supuestos del SEM. Otra limitación es que la matriz de subordinaciones **S** no tiene en cuenta cuántos y cuáles atributos son comunes a los ítems (un ítem está subordinado a otro tanto si comparte un sólo atributo como si comparte todos menos uno). Es decir, el enfoque propuesto intenta validar un aspecto de la relación entre los ítems que es la subordinación, tal como se ha definido, quedando otras posibles relaciones fuera de consideración, esto hace que dicha metodología no sea apropiada para ciertos niveles de la validación.

2.5. Implicaciones del enfoque psicométrico cognitivo

2.5.1. Validez y diseño de pruebas

El desarrollo y correcta utilización de modelos psicométricos es una condición necesaria pero no suficiente para dar cuenta de las competencias o habilidades de los examinados en un proceso completo de evaluación. Los modelos componenciales anteriormente expuestos pueden ser aplicados “a posteriori” en mediciones que no han sido diseñadas para proporcionar información diagnóstica detallada, sin embargo en este tipo de casos suelen ocurrir desde problemas técnicos como la falta de convergencia en la estimación de parámetros hasta problemas más profundos como la no interpretabilidad de los patrones de habilidad resultantes, pues cuando la evaluación no es diseñada con propósitos de diagnóstico es difícil extraer información precisa sobre los atributos correspondientes a la composición cognitiva de dichas mediciones (Rupp y Mislevy, 2007). Por otro lado, los modelos componenciales pueden ser aplicados a mediciones cuyos diseños reflejan el propósito diagnóstico “a priori”, en este caso, las tareas se diseñan de tal manera que permiten obtener información sobre los atributos individuales y sus relaciones. En estos casos los atributos pueden

ser discriminados con facilidad y los resultados de la medición suelen ser interpretados de forma significativa y válida.

Operacionalizar los constructos y sus relaciones es parte esencial del proceso de medición y es por ello que los principios de la teoría psicométrico-cognitiva deben emplearse también en el diseño de tareas e ítems que midan aspectos específicos de las diferencias individuales. Se han presentado dos aproximaciones conceptuales para el diseño de evaluaciones en este contexto, ambas basadas en una perspectiva centrada en el constructo (Messick, 1995). La primera ha sido desarrollada por Embretson (1998) y la segunda por Mislevy, Steinberg y Almond (2003).

La aproximación propuesta por Embretson (1998) incluye relaciones entre el diseño de los test y la validez de constructo. Para Embretson, el impacto potencial de la Psicología Cognitiva en la medición se ha visto disminuido porque los principios cognitivos no se han incorporado en la concepción tradicional de validez de constructo: por una parte, los resultados de la validación de constructo tienen poco impacto en el diseño de los tests ya que, por lo general, se emplean con pruebas ya construidas y se estudian con técnicas como el análisis factorial o mediante correlaciones con otras medidas. Por otra parte, la teoría tiene un rol limitado en el diseño de investigación sobre la validez de constructo tradicional, dependiendo de los resultados empíricos y no a la inversa. Cuando se tienen en cuenta principios cognitivos, la teoría precede y guía el desarrollo de las tareas en constructos específicos.

La autora propone un sistema de diseño cognitivo que incluye un marco conceptual y procedimental de interacción entre el diseño de test y la validez. El marco conceptual distingue dos aspectos de la validez de constructo: la representación de constructo y las redes nomológicas. La representación de constructo se refiere al significado de las puntuaciones de los test y se elabora mediante el entendimiento de los procesos y estrategias empleados por los examinados al resolver los ítems. Para ello se utilizan métodos de la Psicología Cognitiva actual como el modelamiento matemático del ítem y el estudio de las latencias o tiempos de respuesta. En contraste, las redes nomológicas se ocupan de la utilidad del test para medir diferencias individuales. Se estudia mediante la correlación de las puntuaciones del test con otras medidas. La inclusión de estas dos partes en la validez de constructo permite un rol directo de la Psicología Cognitiva, puesto que los ítems son

diseñados para reflejar constructos cognitivos específicos.

El marco procedimental propuesto por Embretson orienta tanto el desarrollo y selección de ítems como la estimación de los parámetros. Los pasos a seguir en un proceso de evaluación cognitiva según este enfoque son: identificar las metas de la medición, identificar características del diseño en el dominio de interés, desarrollar un modelo cognitivo (revisión y selección de la teoría, desarrollo del modelo), evaluar el potencial psicométrico del modelo y probarlo (ajuste al modelo matemático), especificar la distribución de los ítems en su complejidad cognitiva (matriz \mathbf{Q}), generar ítems que cumplan la especificación anterior, evaluar las propiedades psicométricas y cognitivas del dominio del test, ensamblar formas del test y formular hipótesis para contrastar la validez del proceso de medida.

Esta conceptualización de la validez de constructo sugiere que las variables cognitivas o componentes del modelo influyen de forma diferente en la habilidad general y este impacto flexible de las variables cognitivas en la puntuación del test es lo que permite el diseño de test basado en los atributos del modelo, ya que en este contexto los ítems deben ser contruidos para medir aspectos específicos de las diferencias individuales conociendo el impacto o el peso de las variables componenciales. Mediante la investigación en la representación de constructo se pueden diseñar tareas que requieran el dominio ciertas variables cognitivas lo cual puede utilizarse para manipular, seleccionar ítems o generarlos automáticamente. En resumen, Embretson presenta una conceptualización de la validez de la estructura cognitiva que incluye un modelo de relaciones entre tareas y variables cognitivas y su influencia sobre la puntuación del test mediante la combinación de pesos que determinan la naturaleza de la habilidad que mide la puntuación global del test (representación de constructo) y la relación de esta habilidad con variables de criterio externas (red nomológica).

El segundo enfoque, propuesto por Mislevy, Steinberg y Almond (2003), busca establecer un *argumento de evidencia* lo más completo posible sobre las competencias del sujeto examinado, esta aproximación ha sido llamada Diseño Centrado en la Evidencia (DCE) y por medio de ella se vincula el aparato matemático de la Psicometría con la construcción de ítems y el significado de las puntuaciones.

Los autores enfatizan que el argumento de evidencia no puede ser entendido aparte de un contexto y un uso, es por ello que la estructura del argumento incluye aspectos

de la situación, de las acciones del examinado e información sobre historia y contexto de la persona. Las interpretaciones de las acciones son más importantes que las acciones mismas en el proceso de medida, además, la perspectiva psicológica en la que se enmarca determina la naturaleza de cada elemento en el argumento de medida. Para estos autores la estructura del argumento de evidencia y no el modelo psicométrico, debe ser el punto de partida para el diseño de la medición y debe guiar las decisiones conceptuales, analíticas e interpretativas.

En este enfoque se utilizan cuatro modelos: a) el modelo de competencia en el que se especifican y definen las competencias, habilidades y conocimientos que serán evaluados, b) el modelo de evidencia en el que se describe como actualizar las creencias sobre la ejecución de un examinado, con base en la observación de los resultados, c) el modelo de tarea en el que se estructura la situación de medición para obtener los tipos de observaciones que se quiere hacer y d) el modelo de ensamblaje en el que se define el conjunto de tareas que comprenderá la medición.

El enfoque DCE es general y puede ser desarrollado con varias metodologías o empleado en cualquier aplicación componencial, especialmente en mediciones diagnósticas. Mislevy, Steinberg, Almond, Breyer y Johnson (2001) esquematizan la idea de construir modelos de evidencia basados en estructuras recurrentes de relaciones entre las variables del modelo de competencia, estas estructuras recurrentes servirían como *esqueletos de evidencia* alrededor de los cuales se puede crear un número indefinido de tareas individuales. Esto puede ser ventajoso en mediciones que utilizan tareas complejas en las que hay modelos de competencia y observaciones multivariadas que dependen de las variables del modelo de competencia en diferentes combinaciones, y dependencias entre y dentro de las tareas. Las estructuras de relación entre los atributos pueden ser definidas con cierta generalidad y esto hace que las relaciones fundamentales entre las variables del modelo de competencia y el modelo de evidencia puedan ser especificadas y empleadas para construir tareas que superficialmente pueden parecer muy diferentes pero en el fondo poseen una estructura similar. Algunas de las relaciones entre atributos que proponen estos autores y que a continuación se van a presentar ya se han mostrado de forma matemática cuando se han expuesto los diversos modelos componenciales:

1. Relaciones conjuntivas: se requieren múltiples atributos para resolver una tarea y carecer de alguno de ellos causa bajos niveles en la ejecución.
2. Relaciones disyuntivas: se requieren múltiples atributos para resolver la tarea y altos valores en cualquiera de ellos causa altos niveles en la ejecución.
3. Relaciones compensatorias: se requieren múltiples atributos para resolver la tarea que son capturados en una variable observable, altos niveles en cada habilidad implica mayores probabilidades de acertar la tarea, además la carencia en una habilidad es compensada con el alto desempeño en otra.
4. Relaciones de inhibición: se requieren múltiples atributos para resolver la tarea pero hay variables inhibidoras que deben tener bajos valores, una vez este requerimiento se ha cumplido, la ejecución depende principalmente del nivel en los demás atributos.
5. Relaciones de dependencia condicional: se suelen encontrar en evaluaciones de diferentes aspectos de la misma ejecución o entre ítems que tienen material común en el estímulo. Son relaciones entre las variables observables que indican que están relacionadas más allá de lo establecido en el modelo de competencia, de tal forma que la aparición de un atributo condiciona la aparición del otro.

Un tratamiento matemático de las relaciones entre atributos se puede encontrar en Maris (1999) quien expresa algunas de las relaciones anteriormente expuestas en las denominadas *reglas de condensación*.

2.5.2. Establecimiento de vínculos entre Psicología Cognitiva y Psicometría

La Psicología Cognitiva nace formalmente a mediados del Siglo XX con los trabajos de Miller (1956) y Bruner, Goodnow y Austin (1956), aunque sus bases habían sido sentadas mucho antes. Las áreas de investigación actuales son variadas en cuanto a enfoques, métodos y niveles de explicación, aunque comparten dos premisas comunes: la primera es que todas las personas son similares en términos de la maquinaria cognitiva, procesos y limitaciones y la segunda es que la cultura influye en los procesos del pensamiento y aprendizaje. Hay cinco perspectivas teóricas en el estudio de la cognición humana (Greeno, Collins y

Resnick, 1996): diferencial o del rasgo, comportamental, del desarrollo, del procesamiento de información y sociocultural.

La que mayor impacto y uso ha tenido en la Psicometría ha sido la teoría del procesamiento de la información (VanLehn, 1989). Esta perspectiva examina los procedimientos mediante los cuales las personas adquieren, almacenan y utilizan el conocimiento para resolver problemas. También se analizan las reglas, procesos elementales, sistemas de producción y descomposición de tareas para intentar develar las redes conceptuales que subyacen a la ejecución de los sujetos. Los trabajos recientes se han enfocado en el estudio de la influencia del contexto en el procesamiento de la información (Anderson, Greeno, Reder y Simon, 2000). La aplicación de la teoría del procesamiento de la información en tareas de medición ha comenzado en la década de los 70 por investigadores como Carroll y Payne (1976), Hogaobam y Pellegrino (1978) y Sternberg (1977), esta perspectiva ha desembocado en el desarrollo de modelos en los que los atributos son descompuestos en procesos muy finos y dichos modelos se han utilizado para la generación de tutores inteligentes en los que se realiza una retroalimentación inmediata de los errores producidos en el desarrollo de la tarea con la finalidad de corregirlos en el transcurso de la evaluación.

En la rama experimental de la Psicología Cognitiva se han usado modelos de regresión para estimar parámetros que cuantifican el impacto del procesamiento en la dificultad de las tareas, se han modelado las latencias de respuesta, se han programado tutores inteligentes, se han analizado protocolos de respuesta, movimientos oculares y se ha comparado el desempeño de novatos y expertos, entre muchos otros procedimientos cercanos a la Psicometría. Cuando se adapta la teoría cognitiva al ámbito psicométrico ocurre una simplificación de los atributos o componentes, ya que la Psicometría debe operacionalizar los modelos cognitivos para cumplir las metas de la medición, haciendo que los atributos se consideren de una forma mucho más global que en la investigación cognitiva. Para establecer un puente entre los modelos psicométricos y los cognitivos se han desarrollado los modelos componenciales, que se han visto con detalle en la primera sección del presente capítulo, en los que se han unido los modelos matemáticos de procesamiento cognitivo, provenientes principalmente de la rama empírica de la Psicología Cognitiva, con modelos psicométricos como los derivados de la TRI o de las clases latentes.

2.5.3. Características de los atributos

Los atributos en el contexto de la medición psicométrica-cognitiva han sido abordados comúnmente como variables latentes de tipo discreto, pocos autores los han caracterizado como variables categóricas, continuas o logísticas (Von Davier y Yamamoto, 2004; Dimitrov, 2007) a pesar de que algunos teóricos como Roussos, Di Bello, Stout, Hartz, Henson y Templin (2007), Roussos y Karelitz (2004) y Karelitz (2008) consideran que es más factible que las funciones de los atributos tengan un comportamiento logístico a que sean funciones graduadas (con un salto en el punto de corte entre dominio y no dominio) como se supone en la mayoría de los modelos de diagnóstico cognitivo. En otras palabras, en lugar de que las funciones de los atributos estén caracterizadas por una asíntota inferior, una asíntota superior y un punto de corte (o punto de localización), es más plausible dicha función tenga una asíntota inferior, un parámetro de localización (correspondiente a la dificultad del atributo) y un parámetro de pendiente (que correspondería a la discriminación del atributo) llegando a ser muy similar al modelo de 3 parámetros de la TRI (Roussos y Karelitz, 2004). Cuando se asume que los atributos siguen una función logística se pueden estudiar dos importantes características: monotonidad y homogeneidad.

1. Homogeneidad: Cuando las funciones de un atributo o habilidad tienen parámetros de localización o dificultad muy similares en los distintos ítems o tareas que requieren dicha habilidad se dice que los atributos son homogéneos o que presentan lo que Roussos y Karelitz (2004) y Karelitz (2008) denominan *homogeneidad estadística de los atributos*. Por el contrario, cuando los parámetros de localización de las funciones de los atributos varían significativamente a través de los ítems que requieren la habilidad dada se puede decir que presentan heterogeneidad.
2. Monotonidad: Cuando la probabilidad de acertar la tarea aumenta monótonamente con el nivel de habilidad se dice que la función exhibe la propiedad de monotonidad. En el LSDM la monotonidad de la función de los atributos es uno de los criterios de validación (Dimitrov, 2007).

2.6. Modelos cognitivos sobre aritmética

2.6.1. Concepto y evolución de la habilidad aritmética

La cognición matemática se ha convertido en un área de gran interés en las últimas tres décadas. La aritmética es un componente fundamental en la educación de los niños y también en la vida cotidiana adulta. La aritmética básica también constituye un fundamento para habilidades matemáticas avanzadas, resolución de problemas y razonamientos complejos (Seyler, Kirk, & Ashcraft, 2003). En condiciones normales el conteo aparece poco después de que el pequeño comienza a hablar. Entre los 3 y 4 años la mayoría de los niños saben contar correctamente estableciendo una correspondencia entre las palabras numéricas y los objetos que representan (Gallistel y Gelman, 1992). Hacia los 5 años se adquiere el principio de cardinalidad, es decir que el último valor de un conteo representa el número de elementos de un grupo (Bermejo, 1990). Las primeras estrategias aritméticas de los niños se basan en el conteo, primero en voz alta y luego con los dedos u otros medios físicos. En estrategias más refinadas de la adición, como la estrategia del *mínimo*, en la cual el niño (generalmente a partir de los 7 años) frente a un problema como $8 + 3$ empezará por el número mayor y contará a partir de allí. En algún momento de la educación primaria los problemas aritméticos sencillos ya no se resuelven de forma externa, el niño deja de contar los objetos con los dedos o con objetos físicos y empieza a hacerlo mentalmente, en esta línea, llamada aritmética mental, se basan los modelos cognitivos que se expondrán a continuación.

2.6.2. Modelos cognitivos de aritmética mental

En los últimos años, en el marco de la Psicología Cognitiva, se han realizado numerosas investigaciones en la denominada *aritmética mental* que involucra el estudio de las cuatro operaciones aritméticas básicas entre números naturales (suma, resta, multiplicación y división). Estos estudios han profundizado principalmente en las variables que influyen en la complejidad de las tareas de aritmética. Se han usado diversas metodologías para el estudio cognitivo de la aritmética mental entre las que se puede contar el estudio de autoreportes, latencias de respuesta, comparación entre personas con discapacidades neuronales y sujetos sin alteraciones, estudios del sustrato neurológico y comparación entre grupos

de sujetos con diferentes grados de habilidad matemática. Entre las variables estudiadas se puede encontrar una influencia destacada de:

1. Formato de presentación de la ecuación: Yip (2002) ha estudiado los efectos de la forma de presentación de las tareas. Específicamente se ha manipulado la presentación en ecuación normal o revertida (e.g $8 + 4 = ?$ vs $8 = 3 + ?$) encontrando que la forma revertida incrementa la dificultad de la tarea medida en tiempos de respuesta.
2. Resolución de problemas aritméticos redactados: es uno de los tópicos más estudiados por los teóricos de la cognición. Existen numerosas investigaciones y modelos sobre los procesos implicados en este tipo de problemas entre los que se pueden contar los trabajos de Carpenter (1985) y Carpenter y Moser (1984) en los procesos de adquisición y solución de problemas de adición y sustracción y los estudios de Kameenui y Griffin (1989) en las variables implicadas en la solución de problemas verbales. Existe un consenso entre profesores del área de Matemáticas al considerar que los problemas aritméticos expresados en palabras presentan una mayor dificultad que los problemas expresados numéricamente, es por ello que algunos investigadores han estudiado las posibles causas de este fenómeno (Hegarty, Mayer y Monk, 1995). Como en la presente tesis se han diseñado las tareas en formato numérico no se profundizará en estas teorías.
3. Tamaño del número: dentro del área de la aritmética mental han tenido mucha importancia los estudios del efecto del tamaño de los números, en los cuales se analiza el fenómeno de que los problemas compuestos por números grandes se resuelven más despacio y de manera menos precisa que los problemas compuestos por números pequeños (Seyler, Kirk y Ashcraft 2003; LeFevre, Sadesky, y Bisanz, 1996; Penner-Wilger, Leth-Steensen, y LeFevre, 2002). Este efecto se ha evidenciado principalmente en las operaciones de suma y multiplicación, por ejemplo, el tiempo de respuesta y los errores suelen ser mayores para la suma $9 + 8$ que para la suma $2 + 1$. Se han realizado pocos estudios del problema del tamaño en otras operaciones como la sustracción y muchos menos en las operaciones entre números con signo. Seyler y col. (2003) han encontrado una relación fuerte entre los tiempos de respuesta y el tamaño del minuendo en la sustracción simple (especialmente evidente en minuendos

mayores a 11). Estos investigadores concluyen que existe un mayor efecto del tamaño en la sustracción que el que se han encontrado en adición y multiplicación.

4. Tipo de operación y representaciones mentales: Algunos estudios han tenido como objetivo la comparación de la dificultad de las cuatro operaciones aritméticas (suma, resta, multiplicación y división). Dixon, Deets y Bangert (2001) probaron si las personas tenían representaciones similares para las cuatro operaciones, los autores encontraron que las operaciones que se aprenden primero, como la suma, tienen representaciones mentales más fuertes que las operaciones que se aprenden más tarde, como la sustracción o la división. Estos autores sugieren que un factor importante para generar soluciones matemáticas es la representación conceptual de la tarea por parte del examinado, es decir, la representación cualitativa de la relación entre los objetos del problema o tarea. Por lo tanto, la persona puede seleccionar una estrategia o técnica de solución para una operación mental determinada, con base a la relación estructural entre la operación y su representación de la tarea. Conceptualmente, las operaciones que incluyen números negativos son más difíciles de resolver. De acuerdo con la hipótesis filogenética del origen de la comprensión de los números y cantidades, los seres humanos no pueden representar los números negativos de igual manera que los positivos. Cognitivamente se suelen representar mejor los positivos porque no es posible experimentar físicamente o con objetos externos los valores negativos, por lo que se tienden a representar como números *ficticios* (Fischer, 2003). Por su parte, Peterson y Aller (1971) encontraron que las cuatro operaciones aritméticas requieren diferentes tiempos de solución, aunque se presenten números similares en las diferentes tareas. Los autores encontraron que los problemas de adición simple y multiplicación se resuelven más rápidamente que los de sustracción y adición con negativos. Los resultados sugieren que los datos de memoria que requieren las operaciones matemáticas de sustracción y adición con negativos varía en accesibilidad, requiriendo mayores búsquedas de datos en zonas menos accesibles de la memoria.

2.6.3. Aritmética con números enteros

Como ya se ha visto, hay varios estudios del sustrato cognitivo requerido para resolver las operaciones aritméticas básicas con números positivos y aritmética mental, pero son pocos los estudios que se encuentran a nivel de la aritmética con números enteros. Algunos teóricos del aprendizaje de las matemáticas como Fuson (1992) o Schwarz, Kohn y Resnick (1993) reconocen carencias importantes en el estudio de los procesos implicados en la resolución de problemas con números enteros negativos, concretamente en los algoritmos de cálculo. Los estudios realizados hasta el momento se pueden clasificar en tres modelos:

2.6.3.1. Modelo de cuatro estrategias

Campbell (2005) propone el uso de cuatro estrategias para la resolución de este tipo de tareas: a) la *recuperación* en la que el examinado refiere *simplemente recordar la respuesta o que se le viene a la mente*, b) la *procedimental* en la cual el individuo realiza un grupo de procedimientos para resolver la tarea como contar hacia delante, contar hacia atrás o descomponer la ecuación en componentes más simples, c) la basada en *reglas* en la que la persona usa pautas como el inverso aditivo [$a - b = a + (-b)$], el uso del 0 como módulo [$a + 0 = a$, $a - 0 = a$, $a - a = 0$] y las reglas de resolución de paréntesis [$a + (-b) = a - b$, $a - (-b) = a + b$] y d) la *compleja*, compuesta por múltiples pasos en los que se pueden combinar estrategias anteriores. Este autor propone otras estrategias para las operaciones que involucran más de dos dígitos, que no serán abordadas aquí pues no es el objetivo de estudio de la presente tesis.

2.6.3.2. Modelo basado en la teoría de los esquemas

Este modelo es planteado en un estudio de Real, Olea, Ponsoda, Revuelta y Abad (1999) en el que se proponen 10 procesos para resolver tareas de aritmética con números enteros, organizadas en un test llamado *prueba de signos*. Los autores se basan en la definición de esquema de Marshall (1993) que incluye cuatro tipos de conocimientos: a) conocimiento de los rasgos distintivos de un fenómeno o situación, b) conocimiento de las condiciones que los delimitan y de la posibilidad de aplicar determinados procedimientos para resolver problemas, c) conocimientos relativos a la planificación de la solución de un

problema de una categoría dada (estrategias), y d) conocimientos relativos a los algoritmos y reglas de cálculo necesarias para ejecutar el proceso de solución.

Real y Cols. (1999) organizan el contenido de los problemas de la prueba de signos de acuerdo a estos 4 tipos de conocimiento y los transforman en procesos de resolución de cada problema. La operación correspondiente al primer tipo de conocimiento es establecer distinciones entre los cuatro tipos de operación (suma, resta, multiplicación o división), el proceso correspondiente al segundo tipo de conocimiento involucra dos operaciones: evaluar si los términos a y b son positivos o negativos e identificar si el valor absoluto de a es mayor al de b . El tercer proceso es establecer una serie de estrategias que permiten llegar a la solución, dependiendo del modo en que se dispongan los elementos en función de los condicionantes anteriores: cambiar el signo de los términos (a , b) o del resultado (c), cambiar el signo del operador o intercambiar entre sí los términos (a y b). Los autores clasifican en el último tipo de conocimiento la capacidad para llevar a cabo una operación aritmética simple, previa a la obtención de la solución.

Para determinar los procesos de resolución de las tareas los autores determinaron las posibles combinaciones de los elementos de cada problema creando un árbol de decisión en el que se asumen ocho atributos para las operaciones de suma y resta con números enteros, A_1 : adición básica, A_2 : sustracción básica, A_3 : evaluación de que el valor absoluto de a es mayor al de b o viceversa, A_4 : cambio de signo del componente a , A_5 : cambio de signo del componente b , A_6 : cambio de operador suma por resta y viceversa, A_7 : cambio de posición de los miembros de la ecuación y A_8 : cambio del signo al resultado. Los otros dos atributos corresponden a operaciones de multiplicación y división.

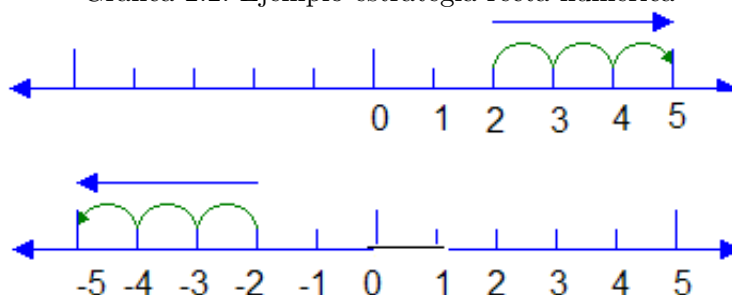
2.6.3.3. Modelos basados en la teoría de enseñanza-aprendizaje

Con base a los modelos de enseñanza de las operaciones aritméticas con números enteros, identificados tanto en la teoría subyacente (Kamii y Joseph, 1989; Maza, 1991, 2001; Miarale, 1984) como en las entrevistas realizadas a los profesores de los colegios objetivo, se pueden definir dos enfoques:

- Enfoque concreto: se trata de una estrategia manipulativa de la enseñanza de la suma y de la resta entre números enteros. La mayoría de estos métodos usan la recta

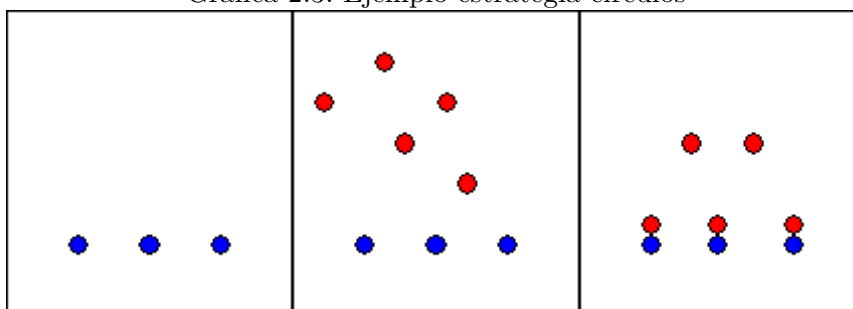
numérica para enseñar a los estudiantes a contar hacia arriba o contar hacia abajo. En el ejemplo de la Gráfica 2.2 se representa la suma $2 + 3 = 5$. Se parte de la representación del 2 en la recta y se aumentan 3 unidades porque el operador es una suma. En el segundo ejemplo se presenta la operación $-2 - 3 = -5$. Se parte de la representación del -2 y se avanza 3 unidades a la izquierda porque el operador es una resta.

Gráfica 2.2: Ejemplo estrategia recta numérica



- Existen otras versiones del enfoque concreto que utilizan otras representaciones para enseñar este tipo de operaciones. Cabe destacar dos aproximaciones: el uso de cargas y de círculos de colores, como en el ejemplo de la Gráfica 2.3 ($3 + (-5) = -2$) en el que se representan los números positivos con círculos azules y los negativos con círculos rojos. Se enseña al alumno que tres de los círculos rojos se *adhieren* a los azules quedando dos de los negativos.

Gráfica 2.3: Ejemplo estrategia círculos



- Enfoque abstracto: Algunos autores usan el concepto *hechos numéricos* para referirse a las operaciones básicas que se enseñan en educación primaria (Maza, 1991; 2001). Siendo a y b números naturales, el *hecho aditivo*, $(a + b)$, se aprende paralelamente

al *hecho sustractivo*, $(a - b, a > b)$. Cuando se han afianzado las operaciones con números naturales se introducen las operaciones con números enteros. El proceso de enseñanza de la suma con enteros comienza con el concepto de orden, con la utilización de la recta y la noción de valor absoluto. Los profesores que siguen una metodología abstracta para enseñar la adición y sustracción con números negativos parten de que el alumno tiene unos conocimientos previos que son: a) la definición de número entero, b) conocimiento de números positivos y negativos, c) concepto de orden entre enteros, d) concepto de valor absoluto y números opuestos y e) concepto de propiedades como inverso aditivo, conmutativa y modulativa. Cuando el alumno maneja dichos conceptos se introduce la suma con números de igual signo con la instrucción: *se suman las cantidades y al resultado se le antepone el mismo signo de los números implicados*, que se plasma en la regla: el resultado de sumar dos positivos siempre es positivo, y el de dos negativos siempre es negativo. La suma con números de diferente signo se enseña con la instrucción: *se toma el valor absoluto de los dos números, se resta el menor del mayor y al resultado se le pone el signo del número mayor*. La solución de los paréntesis se enseña utilizando el concepto de inverso aditivo: *cuando hay paréntesis se debe cambiar el signo de b y cambiar el operador antes de proceder con el resto de la operación*. Por ejemplo: $-8 + (-3) = -8 - 3$ o $-8 - (-3) = -8 + 3$.

Por lo general, los profesores comienzan la enseñanza de las operaciones aritméticas básicas con un enfoque concreto y poco a poco se espera que los alumnos avancen a una comprensión más abstracta del tema mediante la aplicación de reglas. En un estudio de Muldoon, Lewis y Francis (2007) sobre la utilización de estrategias según el tipo de problemas, se concluye que se utilizan más las estrategias concretas en tareas de tipo $-a + b$, que hay una mezcla de estrategias en las operaciones $-a - b$ y que, para resolver tareas de tipo $a - b$, cuando $a < b$, se usan más los enfoques abstractos.

2.7. Modelo cognitivo considerado en la presente tesis

En la presente tesis se parte de un modelo que combina los modelos presentados anteriormente asumiendo un enfoque abstracto y una estrategia basada en la aplicación de reglas. El primer paso para resolver este tipo de tareas se ha denominado *simplificación* y consiste en resolver el paréntesis de acuerdo a dos tipos de reglas: *cambiar el signo de b y cambiar el operador suma por resta* [$a + (-b) = a - b$] o *cambiar el signo de b y el operador resta por suma* [$a - (-b) = a + b$]; estas reglas ya han sido expuestas anteriormente. El test de aritmética que se ha diseñado contiene tareas de tipo $a + (-b)$, $-a + (-b)$, $a - (-b)$ y $-a - (-b)$, con paréntesis que afectan solo al segundo componente de la ecuación (b), de esta manera tras aplicar la regla de simplificación se tienen tareas de tipo $a - b$, $a + b$, $-a - b$ y $-a + b$.

El siguiente proceso que se considera es la *clasificación* de estos cuatro tipos de operación. En este paso, el examinado debe valorar cuales son las reglas apropiadas para aplicar en cada caso. Esta clasificación no es necesaria para los ítems de tipo $a + b$ y $a - b$, cuando $a > b$, pero si para los ítems que involucran suma entre enteros $-a + b$, resta entre enteros $-a - b$ y resta $a - b$ cuando $a < b$. La clasificación errónea de los ítems conducirá a la aplicación equivocada de las reglas. Si la tarea es del tipo $-a + b$ o resta ($a - b$ cuando $a < b$) se usará el tercer atributo que corresponde a la instrucción *se toma el valor absoluto de los dos números, se resta el menor del mayor y al resultado se le pone el signo del número mayor*. Este atributo se ha llamado *operación de enteros 1*.

El cuarto atributo ha sido denominado *operación de enteros 2* e incluye a los ítems que del tipo $-a - b$. Si el ítem ha sido clasificado en esta opción el alumno aplicará la regla: *se suman las cantidades y al resultado se le antepone signo negativo*. Es importante notar que los atributos tres y cuatro son mutuamente excluyentes pues una vez clasificados los ítems se sigue un camino u otro. Los dos atributos restantes que componen la estructura cognitiva que se propone son las operaciones simples de suma y resta ($a > b$) entre números naturales.

Capítulo 3

Análisis de la recuperación de las CCI con el LSDM

3.1. Introducción

Una importante evidencia sobre la validez de los atributos propuestos en la estructura cognitiva es el grado en el que las CCI del modelo de TRI son recuperadas mediante el producto de las probabilidades de los atributos estimadas con el LSDM. La media de las diferencias absolutas entre la CCI original y recuperada es llamada en este contexto MAD (en inglés Mean Absolute Difference) y es calculada mediante la siguiente expresión:

$$MAD_j = \frac{\sum_{m=1}^M |P_{jm} - \hat{P}_{jm}|}{M} \quad (3.1)$$

Donde,

M es el número de niveles de habilidad (θ_i) previamente fijados.

P_{jm} es la probabilidad para el ítem j en el nivel de habilidad m obtenida con la TRI.

\hat{P}_{jm} es la probabilidad para el ítem j en el nivel de habilidad m , estimada por el LSDM.

Idealmente, $MAD = 0$ indica una perfecta recuperación de la CCI. El estadístico MAD permite conocer la calidad del modelo que vincula los ítems con los atributos requeridos para resolverlos, pues valores grandes del estadístico implican que estos vínculos no están bien especificados y por lo tanto se hace necesaria una revisión de los mismos. El objetivo del presente trabajo es obtener y analizar las distribuciones de los valores MAD según

algunas las características de la matriz \mathbf{Q} (cantidad y dificultad de los atributos y número de ítems en el test) y la búsqueda de puntos de corte para dicho estadístico. Los puntos de corte pueden ser usados en la valoración de la recuperación de las CCIs en situaciones aplicadas.

3.2. Procedimiento

3.2.1. Diseño

El presente trabajo se han manipulado tres variables independientes (VI) cuyos niveles se han seleccionado de acuerdo a las características de algunos estudios publicados que usan modelos de diagnóstico y análisis cognitivo: por ejemplo, López y Elosua (2002) han usado siete atributos para explicar la solución de un test de 21 ítems de fracciones; Romero, Ponsoda y Ximénez (2008) han empleado una estructura de seis atributos en 66 ítems de un test de aritmética; Dimitrov y Raykov (2003) han propuesto una estructura de diez procesos para analizar 15 ítems de álgebra, Fischer (1973) ha utilizado una estructura de ocho atributos para descomponer 19 ítems de Matemáticas, Dimitrov (2007) ha empleado una estructura de cinco procesos en un estudio con 20 ítems de comprensión lectora y de la Torre (2008) ha utilizado una estructura de cinco atributos para explicar la ejecución de 15 ítems de resta entre fracciones y números mixtos. Dada la variabilidad en el número de atributos e ítems encontrada en los estudios publicados se ha decidido manipular tres niveles en la VI *número de ítems en el test* (15, 24 y 32), y tres niveles en la VI *número de atributos* (4, 6 y 8). Finalmente, se ha manipulado la dificultad relativa de los atributos, generando atributos ordenados de fácil a difícil siendo A_1 el más fácil y A_4 , A_6 y A_8 los más difíciles en cada condición.

En resumen, en este estudio se ha utilizado un diseño con nueve condiciones ($3 \times 3 = 9$): 4 atributos y 15 ítems (4A15I), 4 atributos y 24 ítems (4A24I), 4 atributos y 32 ítems (4A32I), 6 atributos y 15 ítems (6A15I), 6 atributos y 24 ítems (6A24I), 6 atributos y 32 ítems (6A32I), 8 atributos y 15 ítems (8A15I), 8 atributos y 24 ítems (8A24I) y 8 atributos y 32 ítems (8A32I).

3.2.2. Pasos de la simulación

El procedimiento de simulación consta de cuatro pasos presentados a continuación.

3.2.2.1. Simulación de las matrices \mathbf{Q}

En primer lugar se generaron aleatoriamente nueve matrices \mathbf{Q} ficticias, una para cada condición. En este proceso se aplicaron ciertas restricciones con el fin de obtener matrices exhaustivas. Como las matrices \mathbf{Q} son de carácter binario, en los 9 casos se formaron todas las posibles combinaciones de 0 y 1 excluyendo la combinación de ceros (cada ítem requiere como mínimo un atributo). La matriz de la condición 4A15I incluye las 15 combinaciones binarias posibles ($2^4 - 1 = 15$). Para simular las matrices \mathbf{Q} correspondientes a las condiciones 4A24I y 4A32I ha sido necesario repetir algunas de las 15 combinaciones, para ello se han elegido al azar las filas a repetir adicionándolas a \mathbf{Q} hasta completar el número de ítems deseado. Algunas aplicaciones empíricas de los modelos componenciales (Dimitrov, 2007; López y Elosua, 2002; Romero, Ponsoda y Ximénez, 2006) contienen matrices \mathbf{Q} en la que varios ítems comparten una misma estructura cognitiva de forma similar a las matrices \mathbf{Q} de las condiciones 4A24I y 4A32I.

Para configurar las matrices \mathbf{Q} de las condiciones restantes ha sido necesario emplear un procedimiento de selección de patrones. Este proceso se ha llevado a cabo seleccionando al azar los patrones fila proporcionalmente al número de atributos requerido por el ítem como se expone en la Tabla 3.1 donde NA es el número total de atributos requeridos por el ítem, P es el número de patrones posibles, y cada columna contiene el número de filas seleccionadas para la respectiva condición. Por ejemplo, en las condiciones con 8 atributos, se han incluido mayor cantidad de ítems que requieren de 3 a 6 atributos y menor cantidad de ítems que requieren pocos (1 o 2) o muchos (7 u 8) atributos. Las matrices \mathbf{Q} generadas mediante dicho procedimiento para todas las condiciones se presentan en el Anexo 1 (Capítulo 8).

Tabla 3.1: Número de patrones en **Q** para cada condición

NA	P	6A15I	6A24I	6A32I	NA	P	8A15I	8A24I	8A32I
1	6	1	2	3	1	8	1	2	2
2	15	3	5	7	2	28	1	2	3
3	20	6	9	11	3	56	3	4	6
4	15	3	5	7	4	70	4	7	9
5	6	1	2	3	5	56	3	4	6
6	1	1	1	1	6	28	1	2	3
Total	63	15	24	32	7	8	1	2	2
					8	1	1	1	1
					Total	255	15	24	32

3.2.2.2. Simulación de las respuestas de los sujetos

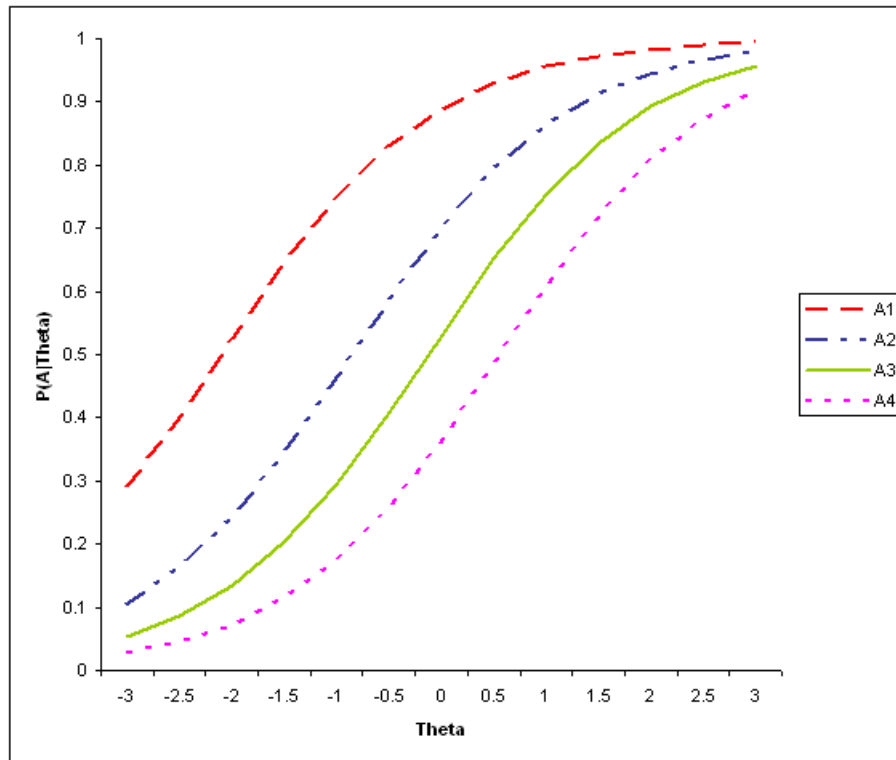
Se han simulado las respuestas de 500 examinados a los ítems de cada condición, de acuerdo a las estructuras cognitivas generadas en el paso anterior. Para ello se ha utilizado el siguiente procedimiento: se generan aleatoriamente tres matrices que contienen los *estados de conocimiento* de los examinados, es decir vectores binarios en los que se simula el dominio o no dominio de los procesos cognitivos de cada examinado, con 4, 6 y 8 atributos respectivamente. En las filas de dicha matriz (que se va a llamar **E**) se representan los individuos y en las columnas los atributos. Los elementos de **E** asumen el valor de 1 si el examinado domina el atributo y 0 si no lo domina. La asignación de valores a esta matriz se ha realizado de forma aleatoria con algunas restricciones mediante las que se busca manipular la dificultad de los atributos.

Al primer atributo se asigna aleatoriamente un 80 % de valores 1 y un 20 % de valores 0 buscando que sea el atributo de menor dificultad relativa, pues se asume que la mayoría de los examinados lo dominan. Los valores del segundo atributo dependerán de los del primero: si $E_{i1} = 0$, entonces $E_{i2} = 1$ con una probabilidad de 0.20 y $E_{i2} = 0$ con una probabilidad de 0.80, por el contrario, si el valor de $E_{i1} = 1$, entonces $E_{i2} = 1$ ó 0 con igual probabilidad ($p = 0.50$), mediante dicho procedimiento se busca que A_2 sea más difícil que A_1 pues un examinado que carece de dominio de A_1 tendrá alta probabilidad de no dominar A_2 . Se procede de forma similar asignando valores binarios a los atributos restantes dependiendo de los valores dados al atributo inmediatamente anterior, de esta manera se obtienen atributos ordenados en dificultad.

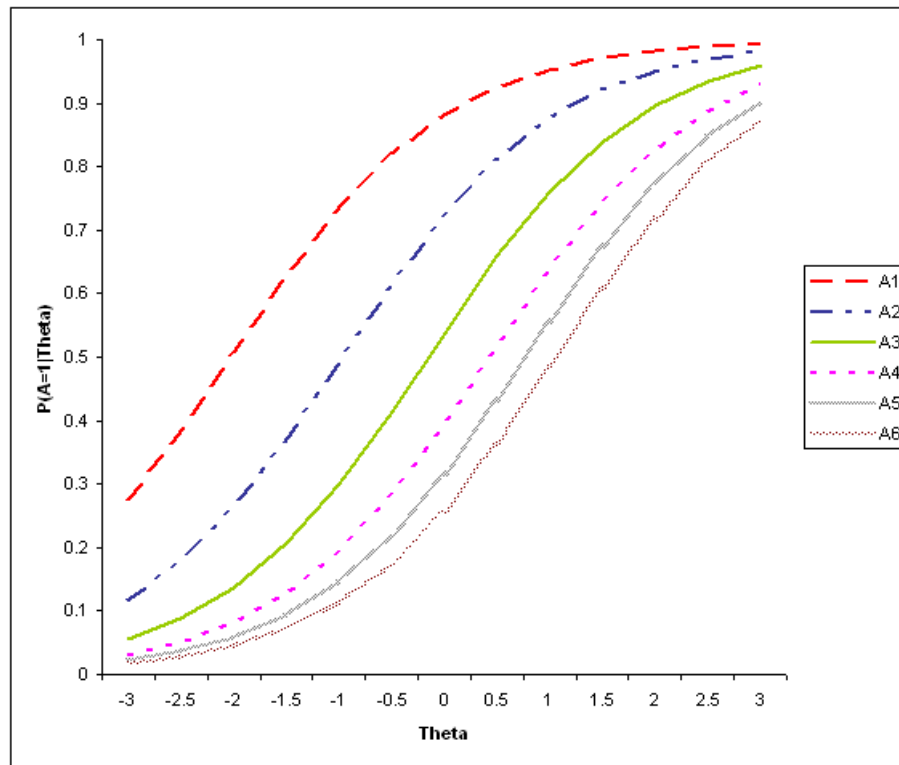
Partiendo del supuesto fundamental del LSDM de que el acierto en el ítem depende

de la conjunción del dominio de los atributos, se multiplican los elementos de la matriz \mathbf{E} de acuerdo a la matriz \mathbf{Q} para obtener la matriz de datos X_{ij} con las respuestas de 500 sujetos a cada uno de los ítems. Por ejemplo, si \mathbf{Q} indica que un ítem particular requiere los atributos A_1 y A_4 la respuesta simulada de un examinado a dicho ítem será acierto solo en el caso en que domine los dos atributos, es decir que los valores correspondientes a estos dos atributos en la matriz \mathbf{E} sea 1, si el examinado careciera de dominio en alguno de los dos (o en los dos) la respuesta asignada será $X_{ij} = 0$. Para verificar la corrección del procedimiento de generación de atributos y su orden creciente de dificultad se han calibrado con el modelo de Rasch las 3 matrices \mathbf{E} y se han representado las CPAs que se obtienen con el LSDM, esto se ha realizado con el objetivo de verificar si se cumple la propiedad de orden de dificultad entre los atributos que se pretendía establecer en la simulación. En las Gráficas 3.1 a 3.3 se puede apreciar que en las tres condiciones cumple el ordenamiento de los atributos en cuanto a su dificultad relativa.

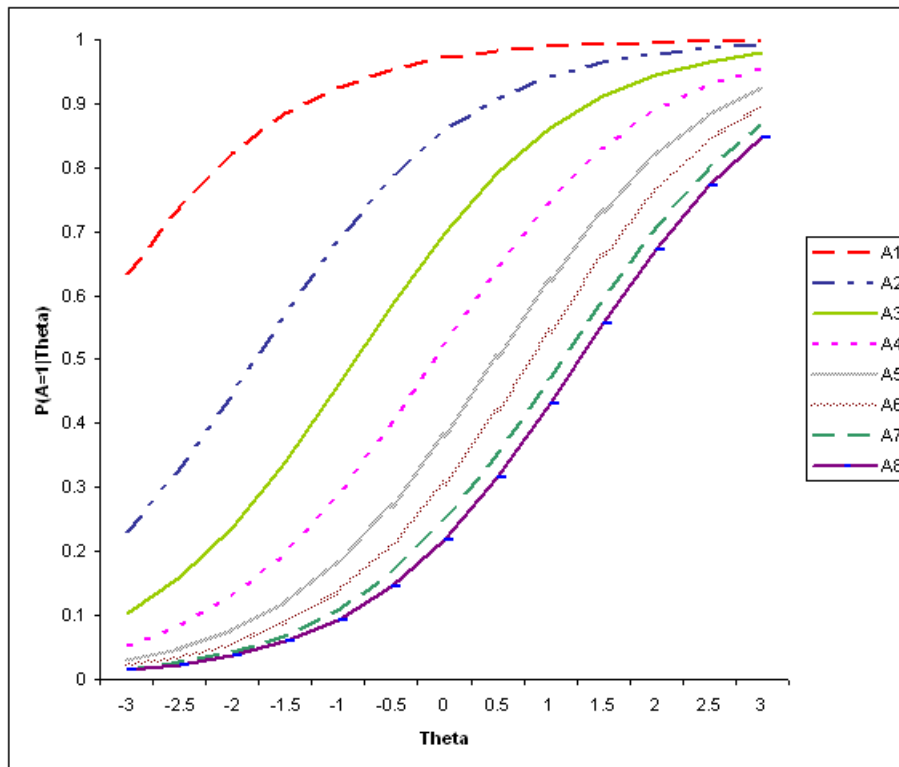
Gráfica 3.1: CPAs de las condiciones con 4 atributos



Gráfica 3.2: CPAs de las condiciones con 6 atributos



Gráfica 3.3: CPAs de las condiciones con 8 atributos



3.2.2.3. Ajuste del modelo TRI

El tercer paso del procedimiento es evaluar el ajuste del modelo de TRI en las matrices de respuesta simuladas (X_{ij}) con el fin de obtener los parámetros de los ítems que serán usados posteriormente en el modelo LSDM. Se ha ajustado el modelo de Rasch con parámetros de discriminación iguales a 1 para todos los ítems. Para realizar la calibración de los ítems y evaluar su ajuste se ha usado el programa *ltm* (Rizopoulos, 2006) que funciona en el entorno R. La estimación de los parámetros en los modelos TRI ha tenido gran importancia en la literatura del área. Usando procedimientos de máxima verosimilitud (ML) se han desarrollado tres métodos de estimación de parámetros llamados condicional, completo y marginal. Una explicación detallada de dichos métodos se presenta en Baker y Kim (2004) y una discusión de las ventajas de cada método se encuentra en Agresti (2002). La otra gran vertiente para la estimación de parámetros es la Bayesiana (Baker y Kim, 2004). El paquete *ltm* que se ha usado para todos los estudios de la presente tesis ajusta los modelos TRI mediante máxima verosimilitud marginal (Marginal Maximum Likelihood Estimation - MMLE).

3.2.2.4. Aplicación del LSDM en las réplicas

Los pasos anteriores se repiten 100 veces para cada condición, obteniendo así un número suficiente de réplicas para aplicar el LSDM en cada una de ellas y obtener la distribución de los valores MAD.

3.3. Resultados

3.3.1. Ajuste del modelo TRI

Las Tablas 3.2 a 3.4 presentan, a modo de ejemplo, los parámetros estimados y el ajuste para cada ítem del modelo de TRI en una réplica seleccionada al azar de las 100 que se han realizado en cada una de las condiciones.

Tabla 3.2: Parámetros, error estándar y ajuste del modelo de Rasch en una réplica de 15 ítems

4A15I				6A15I				8A15I			
b	$s.e$	χ^2	p	b	$s.e$	χ^2	p	b	$s.e$	χ^2	p
-2.46	.145	.079	.779	0.76	.117	1.76	.185	0.86	.116	1.38	.239
-0.94	.120	.565	.452	0.99	.120	1.91	.166	2.44	.160	2.24	.134
-0.33	.119	.578	.447	1.82	.139	3.51	.061	3.53	.225	3.32	.068
-0.01	.121	.140	.709	2.50	.163	2.71	.100	3.44	.218	1.48	.223
-0.68	.119	.570	.450	2.08	.148	0.95	.329	3.09	.194	0.70	.402
0.01	.121	.757	.384	2.33	.157	2.52	.112	3.86	.252	3.29	.070
0.44	.124	.299	.585	2.01	.145	1.14	.285	3.58	.228	3.54	.060
0.28	.122	.563	.453	3.17	.197	2.98	.084	3.20	.200	2.00	.157
0.96	.129	.482	.488	3.45	.214	4.37	.037	3.23	.203	1.51	.218
0.76	.127	.369	.544	3.45	.214	2.84	.092	4.79	.361	2.83	.092
1.31	.133	.703	.402	3.45	.214	4.37	.037	4.46	.316	4.79	.029
1.02	.130	.525	.469	3.58	.222	1.15	.283	3.69	.237	1.17	.278
1.11	.131	.561	.454	3.95	.249	4.49	.034	4.79	.361	1.85	.173
0.46	.124	.612	.434	4.01	.254	1.62	.203	5.25	.435	4.26	.039
1.44	.134	.848	.357	4.51	.299	1.91	.167	5.43	.437	5.43	.020

Tabla 3.3: Parámetros, error estándar y ajuste del modelo de Rasch en una réplica de 24 ítems

4A24I				6A24I				8A24I			
b	$s.e$	χ^2	p	b	$s.e$	χ^2	p	b	$s.e$	χ^2	p
-2.85	.151	-1.62	.184	0.16	.112	0.66	.415	0.62	.113	0.30	.581
-1.19	.126	-0.43	.272	0.78	.118	0.66	.416	0.51	.112	0.32	.569
-0.53	.127	-0.22	.356	1.01	.122	0.31	.572	2.05	.144	0.37	.540
-0.14	.129	-0.10	.110	2.14	.150	0.83	.362	0.98	.118	0.41	.519
-1.19	.126	-0.43	.272	2.08	.148	0.86	.353	3.20	.199	1.56	.211
-0.53	.127	-0.22	.356	2.40	.160	0.87	.349	2.99	.186	0.31	.575
-1.00	.126	-0.33	.221	2.22	.153	1.27	.258	2.25	.151	0.10	.751
-0.13	.129	-0.14	.270	2.57	.167	0.71	.398	1.93	.140	0.65	.418
0.33	.132	0.04	.154	3.17	.197	0.39	.530	4.27	.293	1.11	.291
0.16	.131	0.01	.189	2.47	.163	1.62	.203	3.73	.240	0.79	.374
0.84	.137	0.24	.236	2.65	.170	1.54	.214	3.23	.202	1.62	.202
0.66	.135	0.20	.264	2.97	.186	0.71	.397	3.35	.210	0.91	.339
-0.13	.129	-0.14	.270	3.49	.216	2.22	.135	2.89	.181	0.66	.415
0.33	.132	0.04	.154	3.17	.197	0.77	.378	2.78	.175	0.81	.367
0.16	.131	0.01	.189	4.01	.254	1.32	.250	4.35	.303	1.54	.214
0.33	.132	0.04	.133	3.07	.191	1.62	.203	5.05	.403	1.32	.250
0.98	.138	0.28	.315	2.73	.174	1.43	.231	4.65	.341	1.46	.226
0.98	.138	0.25	.315	3.58	.222	0.45	.499	3.91	.255	1.78	.182
1.20	.140	0.47	.381	3.17	.197	1.76	.184	4.35	.303	1.54	.214
0.33	.132	0.04	.133	3.63	.226	1.42	.233	5.98	.603	2.21	.137
0.98	.138	0.28	.315	4.07	.259	2.18	.139	4.90	.379	2.51	.113
0.98	.138	0.28	.315	4.14	.265	1.03	.309	5.05	.403	2.27	.131
1.35	.141	0.52	.330	4.52	.298	1.16	.280	5.05	.403	2.42	.119
1.35	.141	0.52	.330	4.90	.338	1.42	.232	5.14	.423	6.00	.014

Tabla 3.4: Parámetros, error estándar y ajuste del modelo de Rasch en una réplica de 32 ítems

4A32I				6A32I				8A32I			
<i>b</i>	<i>s.e</i>	χ^2	<i>p</i>	<i>b</i>	<i>s.e</i>	χ^2	<i>p</i>	<i>b</i>	<i>s.e</i>	χ^2	<i>p</i>
-3.06	.155	1.35	.245	-2.04	.136	2.01	.156	0.85	.119	1.30	.253
-1.30	.127	4.78	.029	0.68	.117	0.52	.468	0.67	.117	0.63	.425
-0.55	.128	0.88	.346	0.79	.119	5.15	.023	2.07	.147	0.81	.367
-0.32	.130	1.06	.301	0.84	.120	0.43	.509	2.16	.150	1.06	.302
-3.06	.155	1.35	.245	1.01	.122	0.34	.558	0.95	.120	1.31	.251
-1.30	.127	4.78	.029	1.42	.130	0.83	.360	3.73	.230	0.26	.610
-0.55	.128	0.88	.346	2.20	.152	1.45	.228	3.89	.242	0.18	.665
-0.32	.130	1.06	.301	2.52	.164	1.70	.191	3.15	.193	1.56	.210
-1.11	.127	4.03	.045	1.96	.144	0.74	.387	2.59	.166	0.70	.401
-0.22	.130	1.22	.268	1.89	.142	2.07	.150	2.64	.168	0.44	.503
0.06	.133	1.38	.239	1.60	.134	0.74	.388	2.19	.151	1.02	.312
0.09	.133	1.56	.211	2.22	.153	1.06	.302	4.90	.342	0.84	.359
0.78	.140	0.78	.375	2.16	.151	0.83	.361	4.13	.262	0.54	.461
0.68	.139	2.45	.117	2.31	.156	1.12	.289	4.20	.268	0.40	.526
-1.11	.127	4.03	.045	2.29	.155	0.83	.360	3.68	.227	1.87	.171
-0.22	.130	1.22	.268	2.82	.178	1.50	.220	3.84	.238	0.87	.349
0.06	.133	1.38	.239	2.62	.169	1.70	.192	3.38	.206	0.94	.332
0.09	.133	1.56	.211	2.79	.176	1.64	.200	3.34	.204	0.78	.376
0.78	.140	0.78	.375	3.25	.201	2.55	.110	4.42	.289	1.74	.186
0.68	.139	2.45	.117	3.63	.225	0.61	.432	4.13	.262	1.27	.259
-0.22	.130	1.22	.268	3.59	.222	1.18	.276	3.19	.195	1.55	.213
0.25	.135	1.48	.223	2.88	.181	1.51	.219	5.29	.395	4.86	.027
0.87	.141	0.75	.384	3.49	.216	1.13	.287	4.79	.329	2.81	.093
0.93	.141	3.11	.077	3.41	.210	2.57	.109	5.45	.420	1.38	.239
1.26	.144	0.79	.372	3.54	.219	2.26	.132	4.20	.268	0.87	.349
0.25	.135	1.48	.223	3.63	.225	3.87	.049	5.02	.357	1.38	.239
0.87	.141	0.75	.384	4.35	.281	0.69	.405	4.20	.268	1.75	.186
0.93	.141	3.11	.077	3.95	.248	0.97	.323	4.79	.329	0.80	.370
1.26	.144	0.79	.372	3.73	.232	3.74	.053	5.29	.395	1.59	.207
1.34	.145	0.86	.352	4.60	.303	0.98	.321	5.29	.395	1.64	.200
0.06	.133	1.38	.239	5.00	.344	1.29	.254	5.90	.449	3.08	.079

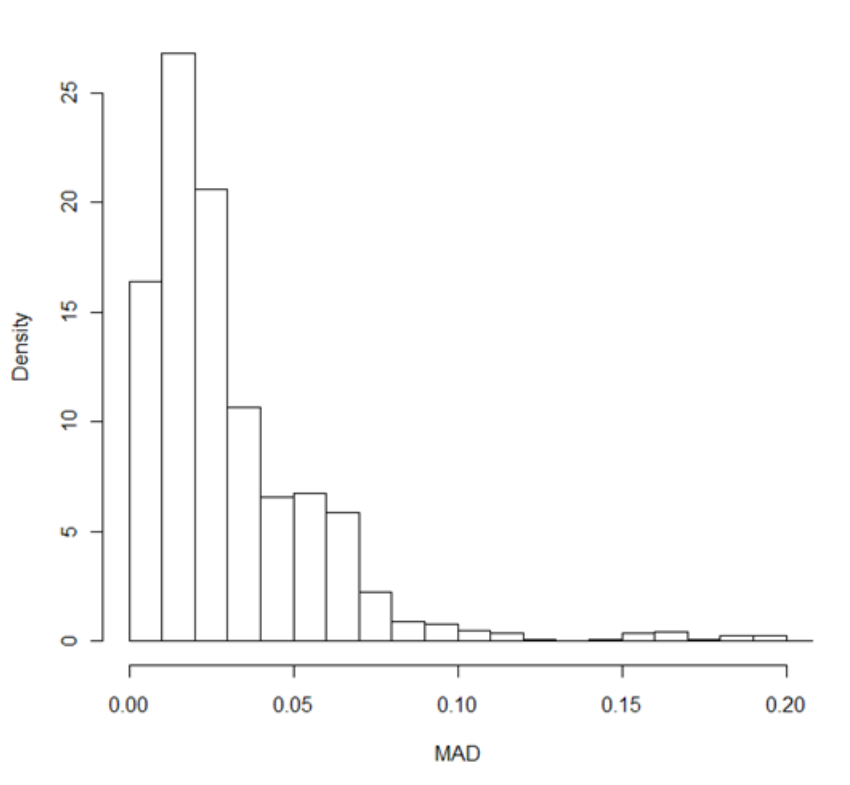
En las Tablas 3.2 a 3.4 se observa que ningún ítem presenta desajuste al modelo de Rasch ya que ningún valor del estadístico χ^2 ha sido significativo ($p < 0.01$). Sin embargo, se aprecia que el ítem que peor ajuste y propiedades presenta es el último pues requiere todos los atributos para ser resuelto y esto ocasiona que sea un ítem muy difícil, con un parámetro b muy alto, y con un error estándar de estimación ($s.e$) también elevado, especialmente en las condiciones con ocho atributos. Es interesante notar que lo anterior no ocurre en la condición 4A15I ya que el último ítem no presenta un parámetro b extremadamente alto ni un error elevado, dicho resultado es explicable pues a medida que

se incrementa el número de atributos la dificultad de los ítems va siendo mayor por las condiciones en que se ha diseñado la simulación. Dentro de cada condición la dificultad de los ítems también se encuentra ordenada de manera incremental pues, según las matrices \mathbf{Q} , los últimos ítems requieren la aplicación de más atributos.

3.3.2. Distribución del estadístico MAD

En la Gráfica 3.4 se presenta la distribución de densidad del estadístico MAD. La distribución se ha trazado con todos los valores MAD simulados, es decir 21300 valores correspondientes a 71 ítems, 3 tamaños de test y 100 réplicas ($71 \times 3 \times 100 = 21300$). En la gráfica se puede observar que la distribución del estadístico es asimétrica.

Gráfica 3.4: Distribución empírica del estadístico MAD



3.3.3. Análisis de la recuperación de las CCI

En primer lugar se estudió la recuperación de cada ítem mediante el análisis de la distribución de los valores MAD trazada con las 100 réplicas en cada una de las 9 condiciones. En la Tabla 3.5, a modo de ejemplo, se presenta la matriz \mathbf{Q} y el percentil 95 de la distribución de los valores MAD para cada ítem en la condición 4A15I. En dicha tabla se puede

observar que los valores del percentil 95 de la distribución disminuyen cuando aumenta el número de atributos que requiere el ítem. En el Anexo 1 (Capítulo 8) se incluyen los resultados correspondientes a las otras condiciones que presentan las mismas tendencias evidenciadas en la Tabla 3.5.

Tabla 3.5: Matriz **Q** y percentil 95 en la condición 4A15I

	A_1	A_2	A_3	A_4	P_{95}
i1	1	0	0	0	.0503
i2	0	1	0	0	.0678
i3	0	0	1	0	.0814
i4	0	0	0	1	.0576
i5	1	1	0	0	.0409
i6	1	0	1	0	.0500
i7	1	0	0	1	.0268
i8	0	1	1	0	.0207
i9	0	1	0	1	.0314
i10	0	0	1	1	.0212
i11	0	1	1	1	.0132
i12	1	0	1	1	.0098
i13	1	1	0	1	.0231
i14	1	1	1	0	.0129
I15	1	1	1	1	.0198

Debido a los resultados anteriores se decidió analizar las distribuciones de valores MAD formando grupos de ítems según el número de atributos que requieren: el grupo 1 (G1) está conformado por los ítems que requieren un atributo, el grupo 2 (G2) por los que requieren 2 atributos y así sucesivamente. En la Tabla 3.6 se presenta la media, mediana y percentil 95 de las distribuciones obtenidas en cada uno de los grupos para cada condición. En dicha tabla se puede observar que, independientemente de la condición, los ítems del G1 (ítems que requieren un solo atributo para su correcta solución) presentan una recuperación mucho mas pobre que los ítems de los demás grupos. También se aprecia que a medida que aumenta el número de atributos requeridos por los ítems disminuyen los valores MAD, siendo mejor la recuperación de las CCI's con el método LSDM. En cuanto al tamaño de la prueba, se observa que la recuperación de las CCI's tiende a ser mejor en las condiciones con menos ítems (15) comparada con los otros dos tamaños de test (24 y 32 ítems).

Tabla 3.6: Estadísticos de la distribución de valores MAD

	4A15I			4A24I			4A32I		
	\bar{x}	Mediana	P_{95}	\bar{x}	Mediana	P_{95}	\bar{x}	Mediana	P_{95}
G1	.0556	.0588	.0640	.0589	.0611	.0692	.0584	.0599	.0719
G2	.0266	.0247	.0450	.0277	.0272	.0490	.0287	.0262	.0517
G3	.0108	.0106	.0199	.0125	.0122	.0200	.0150	.0136	.0241
G4	.0182	.0181	.0198	.0201	.0201	.0230	.0248	.0240	.0249
	6A15I			6A24I			6A32I		
	\bar{x}	Mediana	P_{95}	\bar{x}	Mediana	P_{95}	\bar{x}	Mediana	P_{95}
G1	.0306	.0302	.0381	.0939	.0911	.1158	.0951	.0945	.1181
G2	.0203	.0091	.0523	.0517	.0521	.0717	.0579	.0578	.0890
G3	.0206	.0194	.0442	.0241	.0249	.0439	.0301	.0261	.0697
G4	.0154	.0168	.0256	.0170	.0102	.0471	.0107	.0093	.0217
G5	.0052	.0048	.0093	.0188	.0188	.0258	.0144	.0147	.0204
G6	.0133	.0130	.0153	.0241	.0236	.0306	.0238	.0237	.0277
	8A15I			8A24I			8A32I		
	\bar{x}	Mediana	P_{95}	\bar{x}	Mediana	P_{95}	\bar{x}	Mediana	P_{95}
G1	.0480	.0468	.0603	.1756	.1820	.1964	.1203	.1246	.1672
G2	.0492	.0490	.0549	.0553	.0497	.0862	.0517	.0521	.0646
G3	.0252	.0251	.0449	.0448	.0475	.0674	.0319	.0321	.0462
G4	.0227	.0258	.0379	.0160	.0159	.0261	.0207	.0200	.0368
G5	.0169	.0132	.0331	.0084	.0073	.0159	.0113	.0112	.0223
G6	.0091	.0092	.0132	.0133	.0134	.0189	.0104	.0086	.0208
G7	.0057	.0054	.0085	.0088	.0076	.0159	.0139	.0137	.0184
G8	.0117	.0116	.0140	.0137	.0132	.0172	.0130	.0126	.0162

En la Tabla 3.7 se presentan los puntos de corte para evaluar la calidad de la recuperación de las CCIs de acuerdo a las distribuciones de los valores MAD obtenidas mediante el presente estudio de simulación. Se ha decidido proponer como punto de corte el valor correspondiente al percentil 95 de cada grupo; se ha preferido dicho valor, en lugar de otras medidas como la media o la mediana, por ser un punto más conservador dejando por debajo el 95 % de los valores de la distribución y también porque, como se ha visto, la distribución de los valores MAD es asimétrica. El punto de corte depende del número de atributos requeridos por los ítems ya que el estadístico MAD ha resultado ser sensible principalmente a esta variable. Los valores reportados en la Tabla 3.7 pueden utilizarse como herramientas para la valoración de la calidad de la recuperación de las CCIs en estudios de aplicación con el LSDM. Por ejemplo, si en un estudio se tiene una estructura cognitiva de 6 atributos y un ítem que requiere tres de ellos para ser resuelto correctamente, un valor MAD por encima de 0.051 indicaría una mala recuperación de la CCI y sería necesaria una revisión de los vínculos establecidos en la matriz \mathbf{Q} para dicho ítem.

Tabla 3.7: Puntos de corte para los valores de MAD

	4A	6A	8A
G1	.068	.086	.126
G2	.048	.068	.065
G3	.021	.051	.050
G4	.022	.031	.032
G5		.018	.023
G6		.024	.017
G7			.014
G8			.014

3.4. Conclusiones

En este estudio se ha analizado la recuperación de las CCIs en nueve condiciones en las que se manipula el número de ítems, de atributos y la dificultad relativa de los mismos. Para ello se han simulado las respuestas de 500 sujetos en cada condición, partiendo de la simulación de los patrones de dominio de los atributos (o estados de conocimiento); dichas respuestas se han calibrado con el modelo de Rasch para proceder a aplicar el método LSDM. En primer lugar, se ha trazado la distribución empírica total del estadístico MAD apreciando que la distribución resultante es asimétrica. Posteriormente se ha analizado el percentil 95 de la distribución de los valores MAD, para cada ítem, en cada una de las condiciones. Tras dicho análisis se puede concluir que un aspecto fundamental en la calidad de la recuperación es el número de atributos requerido por los ítems ya que las CCIs se recuperan mejor cuando los ítems requieren, por lo menos, dos atributos. Dicho resultado se hace más evidente cuando aumenta el número de ítems. Es importante aclarar que el estadístico MAD es un promedio de la recuperación en los 13 niveles de habilidad previamente fijados en la escala logística, por lo tanto, los resultados aquí expuestos no son informativos sobre la calidad de la recuperación en niveles de habilidad particulares.

Como el número de atributos requeridos por el ítem ha resultado ser una variable fundamental para el estudio de la recuperación de las CCIs a través de los valores MAD, se ha decidido proponer los puntos de corte siguiendo este criterio. Los puntos de corte pueden ser utilizados por los investigadores en aplicaciones empíricas para evaluar la calidad de los vínculos entre atributos e ítems expresados en la matriz \mathbf{Q} . También pueden ser útiles para decidir y probar nuevas especificaciones en la estructura cognitiva planteada.

Resumiendo, en el presente trabajo se ha estudiado el comportamiento y distribución del estadístico MAD en diversas condiciones de simulación y se han propuesto herramientas y reglas para el uso e interpretación de dicho estadístico en la investigación aplicada.

Capítulo 4

Descripción de los efectos de errores de especificación en la estructura cognitiva

4.1. Introducción

Un elemento clave en el desarrollo y aplicación de los modelos de diagnóstico cognitivo es la correcta especificación de la matriz \mathbf{Q} . Por dicho motivo algunos autores han enfocado su investigación al estudio de los efectos de la introducción de errores de especificación en la estructura cognitiva representada en \mathbf{Q} ; por ejemplo, Rupp y Templin (2008) investigan los efectos de los errores de especificación en la estimación de parámetros y tasas de clasificación con el modelo DINA. En su estudio, generan una matriz \mathbf{Q} con 15 filas que incluye todas las posibles combinaciones binarias que se pueden dar con 4 atributos. Luego, introducen errores de especificación, cambiando elementos de \mathbf{Q} , asegurándose de eliminar ciertas combinaciones de atributos y de representar relaciones de dependencia incorrectas entre componentes. Los resultados de dichos autores muestran que ocurre sobrestimación de los parámetros de desliz (s) cuando se eliminan atributos de la matriz \mathbf{Q} y de los parámetros de adivinación (g) cuando se adicionan atributos. También se reportan tasas altas de error de clasificación para clases de atributos que contienen combinaciones de atributos que han sido eliminados de la matriz \mathbf{Q} .

Baker (1993), estudia la sensibilidad del modelo LLTM a errores de especificación en la matriz de especificación cognitiva (\mathbf{W}). El autor estudia la diferencia entre los parámetros básicos estimados en las condiciones de especificación errónea y los obtenidos en la condición de especificación correcta. Para evaluar dicha diferencia utiliza la raíz media cuadrada de la diferencia (RMSD). En dicho trabajo se consideran seis niveles de error de especificación, alterando el 1 %, 2 %, 3 %, 5 %, 7.5 %, y 10 % de los elementos de la matriz \mathbf{W} , y cuatro tamaños de muestra (20, 50, 100 y 1000). El autor también ha manipulado la cantidad de unos y ceros de los componentes de la matriz \mathbf{W} mediante la generación de dos tipos de matriz: una densa, que contiene gran proporción de unos y otra esparcida, que contiene gran proporción de ceros. Los resultados muestran que los errores introducidos en la matriz \mathbf{W} densa producen valores RMSD más pequeños que los introducidos en la matriz \mathbf{W} esparcida. Aunque el tamaño de la muestra tiene algún efecto sobre los valores RMSD, dicho efecto es muy pequeño comparado con el fuerte efecto del porcentaje de errores introducidos en \mathbf{W} .

De la Torre (2008) propone un método empírico para la validación de estructuras cognitivas con el modelo DINA, llamado método δ . Para estudiar la efectividad del método propuesto, el autor realiza un estudio de simulación introduciendo errores de especificación en una matriz \mathbf{Q} de 5 atributos y 30 ítems. La matriz contiene 10 ítems con un atributo, 10 ítems con dos y 10 ítems con tres. Se ha controlado también que cada atributo sea requerido siempre por 12 ítems. Se analizaron 10 matrices \mathbf{Q} con un vector fila mal especificado, y una matriz \mathbf{Q} con 3 vectores fila erróneos. Los errores de especificación fueron introducidos a los ítems 1, 11 y 21 (que requieren uno, dos y tres atributos respectivamente). Como se esperaba, los parámetros estimados tuvieron poco sesgo cuando se empleaba la matriz \mathbf{Q} verdadera, el índice de discriminación promedio (δ) fue de 0.61. En contraste, los parámetros estimados para los ítems con vectores mal especificados muestran sesgos grandes e índices δ reducidos (en promedio, 0.28). Adicionalmente, algunos ítems correctamente especificados fueron afectados por la manipulación de los vectores de \mathbf{Q} .

El objetivo del presente trabajo es estudiar de manera descriptiva los efectos de la introducción de errores de especificación en la estructura cognitiva (introducidos tanto en la matriz \mathbf{Q} como en el comportamiento de los atributos) sobre los criterios de validación del método de las distancias mínimo-cuadráticas (LSDM). En concreto, en este estudio

se describen los efectos de tres tipos de error introducidos en la matriz \mathbf{Q} . Se ha buscado considerar errores que son comunes dentro del contexto de la medición diagnóstica cognitiva: a) errores en los vínculos ítem-atributo que pueden ocurrir cuando se asume que un ítem requiere un atributo cuando en realidad no o viceversa, b) errores en columnas de \mathbf{Q} causados por la omisión de atributos necesarios o la adición de atributos irrelevantes, y c) errores de dependencia e incompatibilidad entre atributos, que son relaciones entre los atributos no contempladas en la estructura original de \mathbf{Q} . Por otra parte se han estudiado también algunos errores que pueden acontecer por el incumplimiento de supuestos sobre el comportamiento de los atributos como la heterogeneidad y falta de monotonicidad de los atributos (Dimitrov, 2007). Los análisis se han centrado en evaluar el efecto de dichos errores en la recuperación de las CPAs y de las CCI, así como en los valores LSD (residuales del método LSDM).

Para estudiar dichos efectos de la forma mas precisa posible se han generado los datos usando un método de simulación diferente al tradicional puesto que no se generan respuestas de sujetos a ítems para luego calibrarlas con el modelo de la TRI, sino que se simulan directamente las probabilidades de acierto a los ítems en cada nivel fijo de habilidad, es decir, se simula la matriz P_{ij} correspondiente al componente izquierdo de la ecuación básica del LSDM:

$$P_{ij} = \prod_{k=1}^K [P(A = 1|\theta)]^{q_{ik}} \quad (4.1)$$

Este nuevo procedimiento de simulación permite controlar el error generado por: a) el componente aleatorio introducido al simular las respuestas de los sujetos y b) el error de estimación de los parámetros de la TRI, de esta manera se garantiza que los cambios ocurridos en los criterios que se analizan se deben *únicamente* a los errores de especificación introducidos en la matriz \mathbf{Q} o en los atributos cognitivos.

4.2. Método

4.2.1. Procedimiento

4.2.1.1. Errores introducidos en la matriz \mathbf{Q}

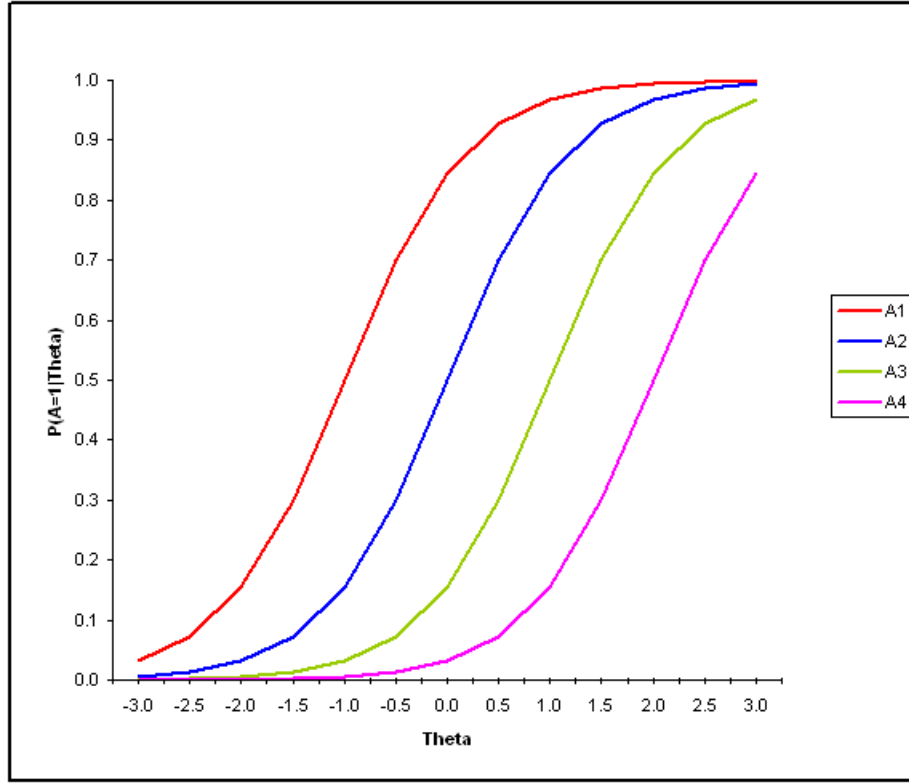
Se parte de una matriz \mathbf{Q} ficticia de 4 atributos y 15 ítems (ver Tabla 4.1), que es la base para la introducción de los errores de especificación. Dicha matriz incluye todas las combinaciones posibles de unos y ceros excepto el vector fila ‘0000’ puesto que cada ítem requiere por lo menos un atributo ($2^4 - 1 = 15$).

Tabla 4.1: Matriz \mathbf{Q} verdadera con 4 atributos y 15 ítems.

	A_1	A_2	A_3	A_4
i1	1	0	0	0
i2	0	1	0	0
i3	0	0	1	0
i4	0	0	0	1
i5	1	1	0	0
i6	1	0	1	0
i7	1	0	0	1
i8	0	1	1	0
i9	0	1	0	1
i10	0	0	1	1
i11	0	1	1	1
i12	1	0	1	1
i13	1	1	0	1
i14	1	1	1	0
i15	1	1	1	1

El primer paso del procedimiento ha sido la generación de una matriz $P_v(A = 1|\theta)$ verdadera que contiene las probabilidades de dominio de los atributos. Para generar dichas probabilidades se ha empleado el modelo TRI de Rasch con parámetros de dificultad: $b_1 = -1$ (A_1), $b_2 = 0$ (A_2), $b_3 = 1$ (A_3) y $b_4 = 2$ (A_4) y 13 niveles de habilidad fijos en la escala logística (-3, -2.5, -2, ... 2, 2.5, 3). Mediante el procedimiento anterior se ha procurado obtener un adecuado comportamiento en los atributos (monotonicidad, la misma discriminación y curvas de atributos que no se crucen entre sí). La Gráfica 4.1 contiene la representación de las CPAs generadas mediante este procedimiento.

Gráfica 4.1: CPAs verdaderas generadas con el modelo de Rasch.



En segundo lugar se ha obtenido la matriz P_{ij} por medio de la multiplicación de las probabilidades de los atributos generadas en el paso anterior, de acuerdo a las especificaciones contenidas en la matriz \mathbf{Q} , como indica la ecuación básica del LSDM:

$$P_{ij} = \prod_{k=1}^K [P_v(A = 1|\theta)]^{q_{ik}} \quad (4.2)$$

En tercer lugar, para evaluar la fiabilidad de este procedimiento, se ha realizado una verificación preliminar en la que se introduce como entrada al análisis con el LSDM la matriz P_{ij} directamente (en lugar de los parámetros de la TRI) y la matriz \mathbf{Q} verdadera (ver Tabla 4.1). Los resultados de esta verificación previa muestran que la matriz de probabilidad de dominio de los atributos estimada por el LSDM es *idéntica* a la matriz verdadera obtenida en el primer paso: $P_{lsdm}(A = 1|\theta) = P_v(A = 1|\theta)$. Por lo tanto, la recuperación de las CCIs de todos los ítems es *perfecta* ($MAD = 0$). Como era de esperarse, los residuales LSD también han sido nulos para todos los niveles de habilidad. Dicha verificación permite asegurar que los cambios que se reportarán en el apartado de resultados, tanto en la forma de las CPAs como en la recuperación de las CCIs, se deben

a los errores introducidos en la matriz \mathbf{Q} y no a alguna imprecisión del método LSDM.

En cuarto lugar se introducen errores de especificación en la matriz \mathbf{Q} de acuerdo a los tres criterios que se exponen a continuación.

Errores en los vínculos ítem-atributo

En este grupo de errores se ha modificado un único valor por cada vector fila (correspondiente a los ítems en \mathbf{Q}), dichos cambios se han introducido solo en aquellos ítems en que ha sido posible, evitando que se genere, por ejemplo, una fila de ceros. En los casos en que se podían hacer dos modificaciones se ha seleccionado al azar la casilla a modificar. Las variables independientes que se han manipulado con sus respectivos niveles son:

1. Tipo de \mathbf{Q} : Densa (D) - Esparcida (S). La matriz \mathbf{Q} densa corresponde a los 5 últimos ítems de la Tabla 4.1. Se denomina densa porque dichos ítems requieren muchos atributos (3 ó 4) para su correcta solución; en otras palabras, tiene muchos unos. La \mathbf{Q} esparcida se compone de los 10 primeros ítems, los cuales requieren pocos atributos (1 ó 2) para su correcta ejecución; en otras palabras, tiene muchos ceros.
2. Dificultad relativa del atributo: Difícil (D) – Fácil (F). Los atributos difíciles son A_3 y A_4 , pues se han simulado con parámetros de dificultad mayores ($b = 1$ y 2 respectivamente). Los atributos fáciles son A_1 y A_2 que se han simulado con parámetros de dificultad menores ($b = -1$ y 0 respectivamente).
3. Tipo de error en el vínculo: Incluir (I) – Excluir (E). La condición *excluír* se refiere a cambiar una entrada de la matriz \mathbf{Q} que originalmente tenía un valor de 1 por un valor 0. Representa la situación en la que un componente requerido por el ítem no es tenido en cuenta en la matriz de especificación cognitiva \mathbf{Q} . La condición *incluir* es inversa pues se refiere a cambiar un valor 0 por 1, representa la situación en la que un componente no requerido por el ítem se vincula a éste erróneamente (ver Tabla 4.2).

El cruce de tres variables con dos niveles cada una produce $2^3 = 8$ condiciones. Las siglas que se utilizarán para nombrar cada matriz \mathbf{Q} con errores de especificación según cada condición son: SFI (esparcida, fácil, incluir), SDI (esparcida, difícil, incluir), SFE

(esparcida, fácil, excluir), SDE (esparcida, difícil, excluir), DFI (densa, fácil, incluir), DDI (densa, difícil, incluir), DFE (densa, fácil, excluir) y DDE (densa, difícil, excluir). Las matrices \mathbf{Q} (tras introducir los errores de especificación) se presentan en las Tablas 4.2 (esparcidas) y 4.3 (densas), se resalta con negrilla cada error introducido y en la última fila se presenta el número total de errores introducidos en cada condición.

Tabla 4.2: Matrices \mathbf{Q} esparcidas con errores de especificación

	SFI				SDI				SFE				SDE			
	A_1	A_2	A_3	A_4	A_1	A_2	A_3	A_4	A_1	A_2	A_3	A_4	A_1	A_2	A_3	A_4
i1	1	1	0	0	1	0	0	1	1	0	0	0	1	0	0	0
i2	1	1	0	0	0	1	0	1	0	1	0	0	0	1	0	0
i3	1	0	1	0	0	0	1	1	0	0	1	0	0	0	1	0
i4	0	1	0	1	0	0	1	1	0	0	0	1	0	0	0	1
i5	1	1	0	0	1	1	1	0	0	1	0	0	1	1	0	0
i6	1	1	1	0	1	0	1	1	0	0	1	0	1	0	0	0
i7	1	1	0	1	1	0	1	1	0	0	0	1	1	0	0	0
i8	1	1	1	0	0	1	1	1	0	0	1	0	0	1	0	0
i9	1	1	0	1	0	1	1	1	0	0	0	1	0	1	0	0
i10	0	1	1	1	0	0	1	1	0	0	1	1	0	0	0	1
	9				9				5				5			

Tabla 4.3: Matrices \mathbf{Q} densas con errores de especificación

	DFI				DDI				DFE				DDE			
	A_1	A_2	A_3	A_4	A_1	A_2	A_3	A_4	A_1	A_2	A_3	A_4	A_1	A_2	A_3	A_4
i1	1	1	1	1	0	1	1	1	0	0	1	1	0	1	0	1
i2	1	1	1	1	1	0	1	1	0	1	1	1	1	0	0	1
i3	1	1	0	1	1	1	1	1	0	1	0	1	1	1	0	0
i4	1	1	1	0	1	1	1	1	1	0	1	0	1	1	0	0
i5	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0
	2				2				5				5			

Errores de adición y fusión de atributos

En este grupo de errores se han adicionado o eliminado atributos por medio de la fusión de columnas de \mathbf{Q} . Las variables independientes que se han manipulado con sus niveles son:

1. Tipo de \mathbf{Q} : Densa (D) - Esparcida (S). Se definen igual que en la condición anterior.
2. Dificultad relativa del atributo: Difícil (D) – Fácil (F). Se definen igual que en la condición anterior.
3. Tipo de error en columna: Adicionar (A) – Fusionar (F). La condición adicionar se refiere a añadir una columna a la matriz \mathbf{Q} y representa la inclusión de atributos

irrelevantes en la matriz de especificación cognitiva. Se ha añadido una columna con elementos binarios aleatorios a cada matriz, respetando la proporción de ítems que requieren el atributo (densa para el tipo D y esparcida para el S). La condición fusionar se refiere a excluir un atributo que en realidad es necesario para resolver las tareas que componen el test, en lugar de eliminarlo se ha decidido fusionarlo con otro atributo.

4. Forma de realizar la fusión: AND – OR. Se refiere a la forma como se realiza la fusión entre los atributos. AND implica que las entradas del atributo fusionado asumirán un valor de 1 si cualquiera de los dos atributos tuviese originalmente un elemento 1. Por el contrario, OR implica que la entrada del atributo fusionado tomará el valor 0 siempre y cuando alguno de los dos atributos tenga un 0 en el elemento correspondiente.

El cruce de cuatro variables, con dos niveles cada una produce $2^4 = 16$ condiciones para un diseño completamente cruzado. Sin embargo en este caso el nivel *añadir* produce solo dos condiciones ya que no se cruza con las variables *tipo de error en columna* ni *dificultad del atributo*. Por esta razón se tienen $2^3 + 2 = 10$ condiciones. Las siglas que se utilizarán para nombrar cada matriz **Q** con errores de especificación de cada condición son: DA (densa, añadir), SA (esparcida, añadir), SFAF (esparcida, fácil, AND, fusionar), SDOF (esparcida, difícil, OR, fusionar), SFAF (esparcida, fácil, AND, fusionar), SDOF (esparcida, difícil, OR, fusionar), DFAF (densa, fácil, AND, fusionar), DDOF (densa, difícil, OR, fusionar), DFAF (densa, fácil, AND, fusionar) y DDOF (densa, difícil, OR, fusionar). Las matrices **Q** tras introducir los errores de especificación se presentan en las Tablas 4.4 (esparcidas) y 4.5 (densas), se resalta con cursiva los casos en los que *no* ha sido posible realizar el cambio por producir vector fila de ceros:

Tabla 4.4: Matrices \mathbf{Q} esparcidas con atributos añadidos o fusionados.

	SA					SFAF			SDAF			SFOF			SDOF		
	A_1	A_2	A_3	A_4	A_5	A_{12}	A_3	A_4	A_1	A_2	A_{34}	A_{12}	A_3	A_4	A_1	A_2	A_{34}
i1	1	0	0	0	1	1	0	0	1	0	0	1	0	0	1	0	0
i2	0	1	0	0	0	1	0	0	0	1	0	1	0	0	0	1	0
i3	0	0	1	0	0	0	1	0	0	0	1	0	1	0	0	0	1
i4	0	0	0	1	1	0	0	1	0	0	1	0	0	1	0	0	1
i5	1	1	0	0	0	1	0	0	1	1	0	1	0	0	1	1	0
i6	1	0	1	0	0	1	1	0	1	0	1	0	1	0	1	0	0
i7	1	0	0	1	1	1	0	1	1	0	1	0	0	1	1	0	0
i8	0	1	1	0	0	1	1	0	0	1	1	0	1	0	0	1	0
i9	0	1	0	1	0	1	0	1	0	1	1	0	0	1	0	1	0
i10	0	0	1	1	1	0	1	1	0	0	1	0	1	1	0	0	1

Tabla 4.5: Matrices \mathbf{Q} densas con atributos añadidos o fusionados

	DA					DFAF			DDAF			DFOF			DDOF		
	A_1	A_2	A_3	A_4	A_5	A_{12}	A_3	A_4	A_1	A_2	A_{34}	A_{12}	A_3	A_4	A_1	A_2	A_{34}
i1	0	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1	1
i2	1	0	1	1	1	1	1	1	1	0	1	0	1	1	1	0	1
i3	1	1	0	1	1	1	0	1	1	1	1	1	0	1	1	1	0
i4	1	1	1	0	1	1	1	0	1	1	1	1	1	0	1	1	0
i5	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1

Errores de dependencia secuencial e incompatibilidad entre atributos

En este grupo de errores se han modificado componentes de \mathbf{Q} para generar relaciones entre pares de atributos que no estaban contempladas en la \mathbf{Q} original. La dependencia secuencial ocurre cuando un atributo que es requerido para resolver un ítem siempre está precedido de otro, en otras palabras el conjunto de ítems que requieren un atributo es subconjunto propio del conjunto de ítems que requieren el otro atributo. Las relaciones de incompatibilidad son todo lo contrario: cuando un atributo se requiere para resolver un ítem el otro no puede estar presente. Las variables independientes que se han manipulado en este caso son:

1. Tipo de \mathbf{Q} : Densa (D) - Esparcida (S). Se definen igual que en las condiciones anteriores.
2. Dificultad del atributo: Difícil (D) – Fácil (F). Se definen igual que en las condiciones anteriores.
3. Tipo de relación entre atributos: Dependencia secuencial simple (R1), dependencia secuencial compleja (R2) e Incompatibilidad (I). Al igual que en el caso anterior no se realizaron cambios cuando se produce un vector fila de ceros. En la condición

R1 se realizaron los cambios necesarios para generar dependencia secuencial entre parejas de atributos (A_1, A_2 y A_3, A_4). Por ejemplo, siempre que un ítem requiere la aplicación A_2 también debe estar presente A_1 (véase las condiciones SFR1 y DFR1 de las Tablas 4.6 y 4.7). En la condición I se definen relaciones de incompatibilidad entre pares de atributos (A_1, A_2 y A_3, A_4) de tal forma que cuando un atributo se requiere para resolver un ítem el otro atributo no puede estar presente (véase Tablas 4.8 y 4.9). En la condición R2 se establecen relaciones de dependencia entre los cuatro atributos, de tal forma que A_3 y A_4 sean secuencialmente dependientes de A_2 , es decir si $q_{i3} = 1$ o $q_{i4} = 1$, entonces, $q_{i2} = 1$; además, A_2 también se hace dependiente de A_1 , por lo tanto, si $q_{i2} = 1$, entonces, $q_{i1} = 1$ (ver tablas 4.6 y 4.7).

Este caso contempla tres variables independientes, dos de ellas con 2 niveles y una con 3 niveles lo que produciría $2 \times 2 \times 3 = 12$ condiciones. Sin embargo, sólo se analizaron 10 condiciones porque el nivel de dependencia secuencial compleja tiene únicamente dos combinaciones SR2 (esparcida, compleja) y DR2 (densa, compleja) pues dicha condición implica dependencias entre todos los atributos y no solo entre parejas de atributos. Las siglas para cada condición son: SFR1 (esparcida, fácil, dependencia simple), SDR1 (esparcida, difícil, dependencia simple), DFR1 (densa, fácil, dependencia simple), DDR1 (densa, difícil, dependencia simple), SFI (esparcida, fácil, incompatible), SDI (esparcida, difícil, incompatible), DFI (densa, fácil, incompatible) y DDI (densa, difícil, incompatible). En las Tablas 4.6 a 4.9 se presentan las matrices \mathbf{Q} esparcidas y densas con los errores de dependencia e incompatibilidad, en negrilla se identifican los cambios que se han introducido y en la última fila se presenta el número total de modificaciones.

Tabla 4.6: Matrices **Q** esparcidas con errores de dependencia

	SFR1				SDR1				SR2			
	A ₁	A ₂	A ₃	A ₄	A ₁	A ₂	A ₃	A ₄	A ₁	A ₂	A ₃	A ₄
i1	1	0	0	0	1	0	0	0	1	0	0	0
i2	1	1	0	0	0	1	0	0	1	1	0	0
i3	0	0	1	0	0	0	1	0	1	1	1	0
i4	0	0	0	1	0	0	1	1	1	1	0	1
i5	1	1	0	0	1	1	0	0	1	1	0	0
i6	1	0	1	0	1	0	1	0	1	1	1	0
i7	1	0	0	1	1	0	1	1	1	1	0	1
i8	1	1	1	0	0	1	1	0	1	1	1	0
i9	1	1	0	1	0	1	1	1	1	1	0	1
i10	0	0	1	1	0	0	1	1	1	1	1	1
	3				3				10			

Tabla 4.7: Matrices **Q** densas con errores de dependencia

	DFR1				DDR1				DR2			
	A1	A2	A3	A4	A1	A2	A3	A4	A1	A2	A3	A4
i1	1	1	1	1	0	1	1	1	1	1	1	1
i2	1	0	1	1	1	0	1	1	1	1	1	1
i3	1	1	0	1	1	1	1	1	1	1	0	1
i4	1	1	1	0	1	1	1	0	1	1	1	0
i5	1	1	1	1	1	1	1	1	1	1	1	1
	1				1				2			

Tabla 4.8: Matrices **Q** esparcidas con errores de incompatibilidad

	SFI				SDI			
	A ₁	A ₂	A ₃	A ₄	A ₁	A ₂	A ₃	A ₄
i1	1	0	0	0	1	0	0	0
i2	0	1	0	0	0	1	0	0
i3	0	0	1	0	0	0	1	0
i4	0	0	0	1	0	0	0	1
i5	1	0	0	0	1	1	0	0
i6	1	0	1	0	1	0	1	0
i7	1	0	0	1	1	0	0	1
i8	0	1	1	0	0	1	1	0
i9	0	1	0	1	0	1	0	1
i10	0	0	1	1	0	0	1	0
	1				1			

Tabla 4.9: Matrices **Q** densas con errores de incompatibilidad

	DFI				DDI			
	A1	A2	A3	A4	A1	A2	A3	A4
i1	0	1	1	1	0	1	1	0
i2	1	0	1	1	1	0	1	0
i3	1	0	0	1	1	1	0	1
i4	1	0	1	0	1	1	1	0
i5	1	0	1	1	1	1	1	0
	3				3			

El último paso del procedimiento ha sido la aplicación del modelo LSDM y la obtención de los criterios de interés: MAD, LSD y CPAs. Cabe recordar que el LSDM ha sido aplicado usando como entrada directamente la matriz de probabilidad de acierto a los ítems en los niveles fijos de habilidad (P_{ij}) en lugar de emplear los parámetros de la TRI, esto se ha hecho con el objeto de evaluar con mayor pureza los efectos de los errores de especificación introducidos en **Q**.

4.2.1.2. Errores de heterogeneidad y falta de monotonicidad de los atributos

En este escenario **no** se han introducido errores de especificación en la matriz **Q**, lo que se ha alterado es la forma de generar la matriz de probabilidad de los atributos $P(A = 1|\theta)$ para conseguir atributos que no cumplan las propiedades esperadas, en otras palabras, que sean heterogéneos y no-monótonos. Al modificar las CPAs cambian también las probabilidades de respuesta a los ítems P_{ij} . En este caso se han manipulado dos variables independientes con cuatro niveles cada una:

- Monotonicidad de los atributos: muy buena - buena - pobre - muy pobre. Cuando la probabilidad de dominio de un atributo se incrementa de forma monótona con el nivel de habilidad se puede decir que la función exhibe la propiedad de monotonicidad. Para manipular esta variable se han generado las probabilidades de dominio de los atributos siguiendo una función cúbica en lugar de usar el modelo TRI de Rasch como en las condiciones precedentes. La función con la que se han generado es:

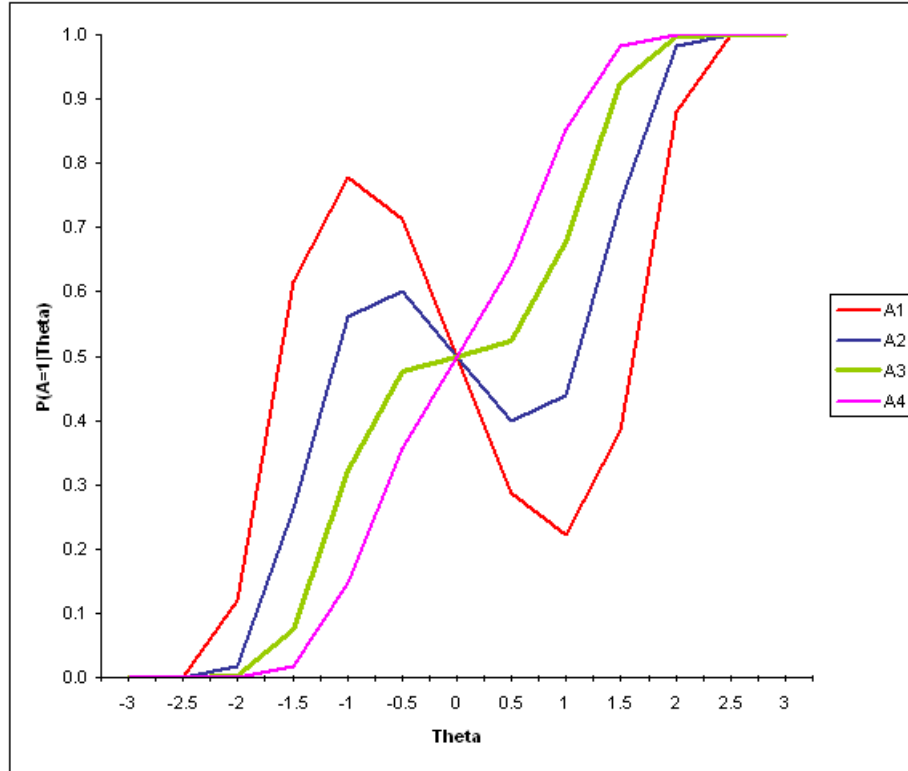
$$P_v(A = 1|\theta) = b_i\theta + c_i\theta^3 \quad (4.3)$$

Donde,

$$b_{A_1} = -2, b_{A_2} = -1, b_{A_3} = 0, b_{A_4} = 1, c = 0.75$$

En la Gráfica 4.2 se puede observar que cuando se usa la función cúbica para generar las CPAs, éstas se ordenan en grado de monotonicidad: A_1 presenta una monotonicidad muy pobre, A_2 pobre, A_3 buena y A_4 muy buena.

Gráfica 4.2: CPAs para las condiciones de falta de monotonicidad



- Heterogeneidad de los atributos: muy heterogéneos - heterogéneos - homogéneos - muy homogéneos. Cuando las funciones de los atributos presentan dificultades iguales en los diferentes ítems que requieren dicha habilidad se dice que los atributos tienen homogeneidad estadística (Roussos y Karelitz, 2004). Por el contrario, cuando la dificultad de los atributos varía significativamente entre los ítems que los requieren se puede decir que los atributos son heterogéneos. Para generar las CPAs de esta condición se ha usado un parámetro b diferente para cada ítem (cada atributo tiene una dificultad diferente según el ítem que lo requiere). Los parámetros b han sido seleccionados al azar de una distribución normal con los siguientes parámetros: $A_1 \sim N(-1, 0.7)$, $A_2 \sim N(0, 0.5)$, $A_3 \sim N(1, 0.3)$ y $A_4 \sim N(2, 0.05)$. Los parámetros utilizados se pueden ver en la Tabla 4.10.

Tabla 4.10: Parametros b del atributo en cada ítem para la condición de heterogeneidad

	A_1	A_2	A_3	A_4
i1	-0.223	-0.465	0.802	1.947
i2	-0.617	-0.103	0.814	1.830
i3	-0.786	-0.149	0.459	1.907
i4	-0.277	-0.253	1.343	2.021
i5	-1.603	0.688	1.192	1.996
i6	-0.501	0.781	0.614	2.045
i7	-0.516	0.474	0.902	1.987
i8	-0.688	-0.386	0.865	2.009
i9	-0.464	0.070	0.662	1.997
i10	-0.535	0.003	0.824	2.000
i11	-0.027	0.501	1.649	2.007
i12	-0.216	0.202	0.922	1.990
i13	-2.212	-0.657	0.629	2.147
i14	-0.636	-0.021	0.396	1.980
i15	0.090	0.201	1.562	2.046

Para obtener la matriz P_{ij} se procedió como en las condiciones precedentes, si bien, los parámetros de la CPA de cada atributo son los indicados en la tabla anterior para cada ítem. De acuerdo con este procedimiento A_1 es el atributo más heterogéneo (parámetros b generados con una desviación estándar mayor) y A_4 el más homogéneo (parámetros b generados con una desviación estándar menor), en otras palabras los atributos se han simulado con una homogeneidad creciente.

El último paso de este procedimiento ha sido nuevamente la aplicación del modelo LSDM y la obtención de los criterios de interés: MAD, LSD y CPAs. Cabe recordar que el LSDM ha sido aplicado usando como entrada directamente la matriz de probabilidad de acierto a los ítems en los niveles fijos de habilidad (P_{ij}) obtenida tanto en la condición de heterogeneidad como en la de no-monotonidad (en lugar de emplear los parámetros de la TRI) y se ha comparado, en el caso de la heterogeneidad, con las CPAs producidas con la P_{ij} original en la que los atributos tienen la misma dificultad a lo largo de todos los ítems [$b_1 = -1$ (A_1), $b_2 = 0$ (A_2), $b_3 = 1$ (A_3) y $b_4 = 2$ (A_4)].

4.2.2. Métodos de análisis estadístico

Las variables dependientes que han sido objeto de estudio en este trabajo son los valores LSD y MAD. También se han valorado los efectos de los errores introducidos en \mathbf{Q} en la estimación de las CPAs mediante: a) una descripción gráfica de las CPAs estimadas y

sus diferencias con las verdaderas, b) el análisis del comportamiento lógico de las CPAs (monotonidad y dificultad relativa) y c) la raíz del error cuadrado medio (RMSE) de la recuperación de la matriz $P_v(A = 1|\theta)$ calculado con la siguiente ecuación:

$$RMSE = \sqrt{\frac{1}{M} \sum_{m=1}^M (P_{lsdm}(A = 1|\theta) - P_v(A = 1|\theta))^2} \quad (4.4)$$

donde,

M corresponde a los 13 niveles fijos de habilidad.

$P_{lsdm}(A = 1|\theta)$ son las probabilidades de dominio de los atributos estimadas con el LSDM.

$P_v(A = 1|\theta)$ son las probabilidades *verdaderas* de dominio de los atributos.

Para valorar la recuperación de las CCIs del presente estudio se han utilizado los puntos de corte hallados en un estudio previo (véase Capítulo 3) para una condición de 4 atributos y 15 ítems como la que nos ocupa. Dichos valores son: 0.068 cuando el ítem requiere un atributo, 0.048 cuando requiere dos atributos y 0.021 cuando requiere 3 o 4 atributos (ver Tabla 3.7 del Capítulo 3).

4.3. Resultados

4.3.1. Errores de especificación en la matriz \mathbf{Q}

4.3.1.1. Errores en los vínculos ítem-atributo

Análisis de los efectos en la recuperación de las CCIs

La Tabla 4.11 presenta los valores MAD para evaluar la recuperación de las CCIs. En dicha tabla se han señalado con un asterisco los ítems mal recuperados, cuyos valores MAD exceden el punto de corte establecido y con cursiva los ítems que **no** han sido objeto de modificación alguna. El número de atributos que se ha tenido en cuenta para la valoración de la recuperación son los que requieren los ítems tras la introducción de errores en \mathbf{Q} .

Tabla 4.11: Estadísticos MAD para evaluar la recuperación de la CCI.

	SFI	SDI	SFE	SDE	DFE	DDE
i1	0.136*	0.150*	0.000	0.262*	0.000	0.000
i2	0.018	0.056*	0.019	0.155*	0.002	0.000
i3	0.006	0.118*	0.011	0.000	0.000	0.048*
i4	0.002	0.029*	0.003	0.018	0.000	0.071*
i5	0.050*	0.112*	0.015	0.203*	0.000	0.000
i6	0.012	0.072*	0.002	0.055		
i7	0.003	0.031*	0.001	0.190*		
i8	0.012	0.085*	0.023	0.033		
i9	0.004	0.037*	0.007	0.152*		
i10	0.000	0.061 *	0.003	0.013		

* ítems mal recuperados según el punto de corte.

No ha sido posible obtener resultados para las condiciones de inclusión con la matriz **Q** densa (DDI y DFI) porque al introducir los errores de especificación dichas matrices resultaban singulares.

Los resultados presentados en la Tabla 4.11 indican que la introducción de errores de especificación en los vínculos ítem-atributo produce un empeoramiento de la recuperación de las CCIs con respecto a la condición base (**Q** verdadera); recuérdese que en dicha condición la recuperación de CCIs fue perfecta (valores MAD igual a 0 para todos los ítems). También se puede observar, en las condiciones con matriz esparcida, que la recuperación de los ítems que no han recibido alteración no ha sido mejor que la de los ítems que contienen errores de especificación.

En la condición SFI la media de los valores MAD en los ítems sin cambio ha sido 0.05 y con cambio 0.02; en la condición SDI la media de los ítems sin cambio ha sido 0.061 y con cambio 0.075, en la condición SFE la media de los ítems sin cambio y con cambio ha sido 0.008 y en la condición SDE las medias han sido 0.12 (con cambio) y 0.08 (sin cambio). De los resultados anteriores se puede deducir que el LSDM permite conocer que existen errores de especificación pero no aporta información concreta sobre cuál es el tipo de ítem (con error o no) en el que ocurre la mala especificación.

La recuperación es evidentemente peor cuando los errores se introducen en los atributos difíciles (SDI, SDE, DDE), especialmente cuando el error implica incluir atributos difíciles en ítems que originalmente no lo requerían (condición SDI) ya que en dicha condición todos los ítems han resultado significativamente mal recuperados. La mitad de los ítems de la

condición SDE también están inadecuadamente recuperados. Por otra parte, los errores que involucran atributos fáciles no han tenido efecto sobre la recuperación de las CCIs pues en la condición SFE (esparcida, fácil, excluir) todos los ítems han sido bien recuperados a pesar de tener errores de especificación. En la condición SFI (esparcida, fácil, incluir) solo dos ítems (i1 e i5) presentan una pobre recuperación y en la condición DFE todos han sido correctamente recuperados.

Comparativamente con las matrices esparcidas, los cambios en las matrices densas han tenido un menor efecto. Esto puede deberse a que las matrices densas han recibido un sólo cambio que afecta poco a los ítems.

Análisis de los valores residuales (LSD)

En la Tabla 4.12 se presentan los valores LSD para cada condición, en dicha tabla se puede observar que dichos residuales se incrementan significativamente en las condiciones que implican cambiar un valor 1 por 0 (exclusión), siendo aún más altos en las condiciones en las que se modifica la matriz **Q** esparcida. El comportamiento de los valores LSD es similar al encontrado en otros estudios con el LSDM: sus valores son más altos en los niveles bajos de habilidad y disminuyen progresivamente conforme la habilidad va en aumento. Los resultados expuestos parecen indicar que los valores LSD podrían ser una herramienta importante para la detección de errores de especificación en la matriz **Q** pues sufren un incremento importante con respecto a la condición de base en la que todos sus valores son iguales a 0.

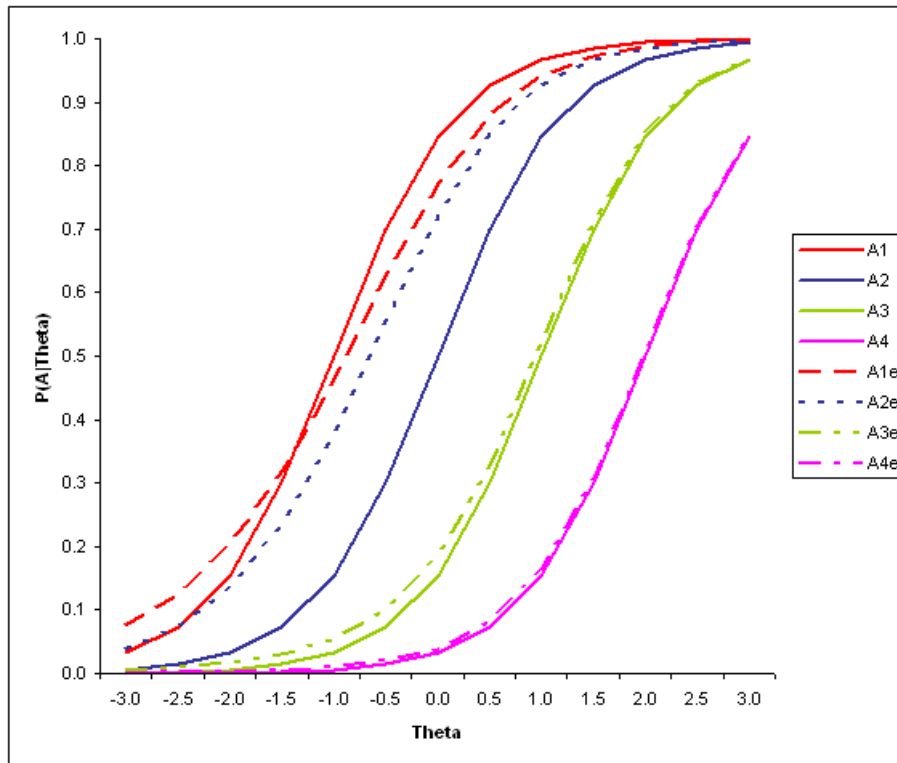
Tabla 4.12: Valores residuales LSD en cada nivel fijo de habilidad

θ	SFI	SDI	SFE	SDE	DFE	DDE
-3	0.440	0.668	0.724	1.147	0.236	0.240
-2.5	0.372	0.603	0.588	1.004	0.231	0.240
-2	0.308	0.547	0.459	0.863	0.221	0.239
-1.5	0.246	0.499	0.339	0.724	0.200	0.238
-1	0.185	0.450	0.235	0.588	0.166	0.236
-0.5	0.126	0.397	0.150	0.458	0.120	0.231
0	0.075	0.339	0.085	0.339	0.074	0.221
0.5	0.040	0.275	0.043	0.235	0.039	0.200
1.0	0.018	0.206	0.020	0.150	0.019	0.166
1.5	0.008	0.139	0.009	0.085	0.008	0.120
2	0.004	0.082	0.004	0.043	0.003	0.074
2.5	0.001	0.043	0.001	0.021	0.002	0.039
3	0.001	0.020	0.001	0.001	0.001	0.019

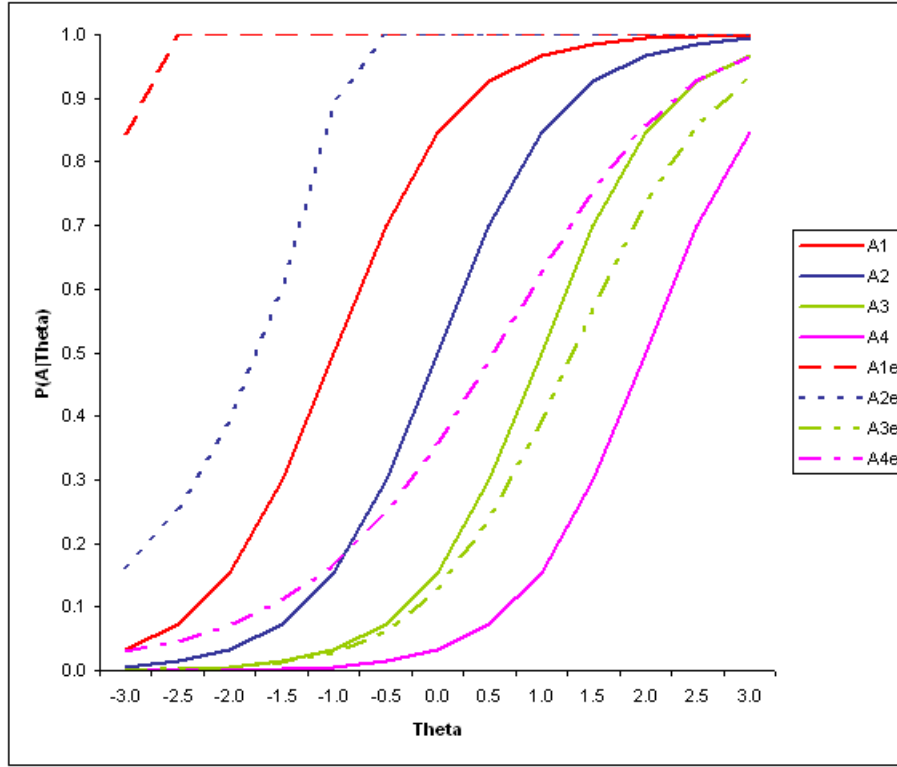
Análisis de los efectos sobre las CPAs

En las Gráficas 4.3 y 4.4 se presentan las CPAs estimadas con el LSDM tras la *inclusión* de vínculos en la matrices esparcidas (SFI y SDI respectivamente) y se comparan con las CPAs originales.

Gráfica 4.3: CPAs condición esparcida, fácil, incluir (SFI)



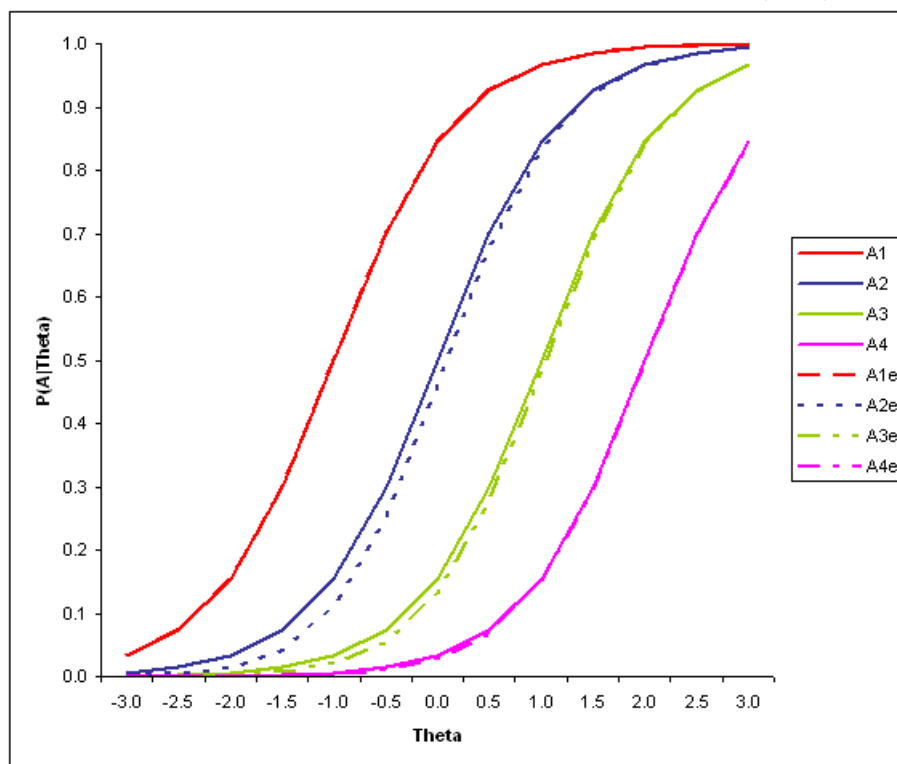
Gráfica 4.4: CPAs condición esparcida, difícil, incluir (SDI)



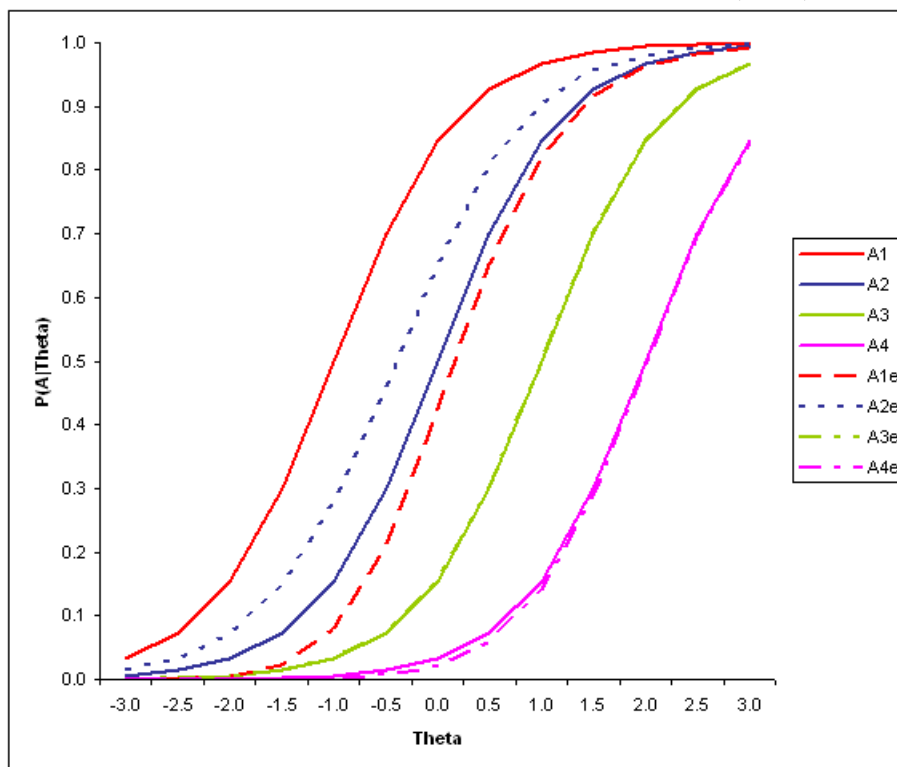
Si se comparan los resultados de las dos condiciones se puede observar que las CPAs se recuperan peor cuando se introducen errores en los atributos difíciles (condición SDI). Contrariamente a lo que se esperaba, las curvas más afectadas son las correspondientes a los atributos fáciles, que no han sido modificados (ver Gráfica 4.4). El LSDM tiende a sobreestimar las probabilidades de estos dos atributos, especialmente en el caso de A_1 , cuya probabilidad pasa a ser 1 en casi todos los niveles de habilidad. Los valores RMSE de dicha condición han sido: 0.488 (A_1), 0.381 (A_2), 0.064 (A_3) y 0.276 (A_4). En cuanto a la inclusión de vínculos en atributos fáciles (SFI) se observa poca afectación en las curvas, excepto para el atributo A_2 , cuyas probabilidades se han sobreestimado (ver Gráfica 4.3), esto se refleja en los valores RMSE de 0.042 (A_1), 0.234 (A_2), 0.017 (A_3) y 0.005 (A_4).

Las Gráficas 4.5 y 4.6 representan las condiciones de *exclusión* de vínculos en los atributos fáciles de las matrices esparcida y densa respectivamente (condiciones SFE y DFE).

Gráfica 4.5: CPAs condición esparcida, fácil, excluir (SFE)



Gráfica 4.6: CPAs condición densa, fácil, excluir (DFE)

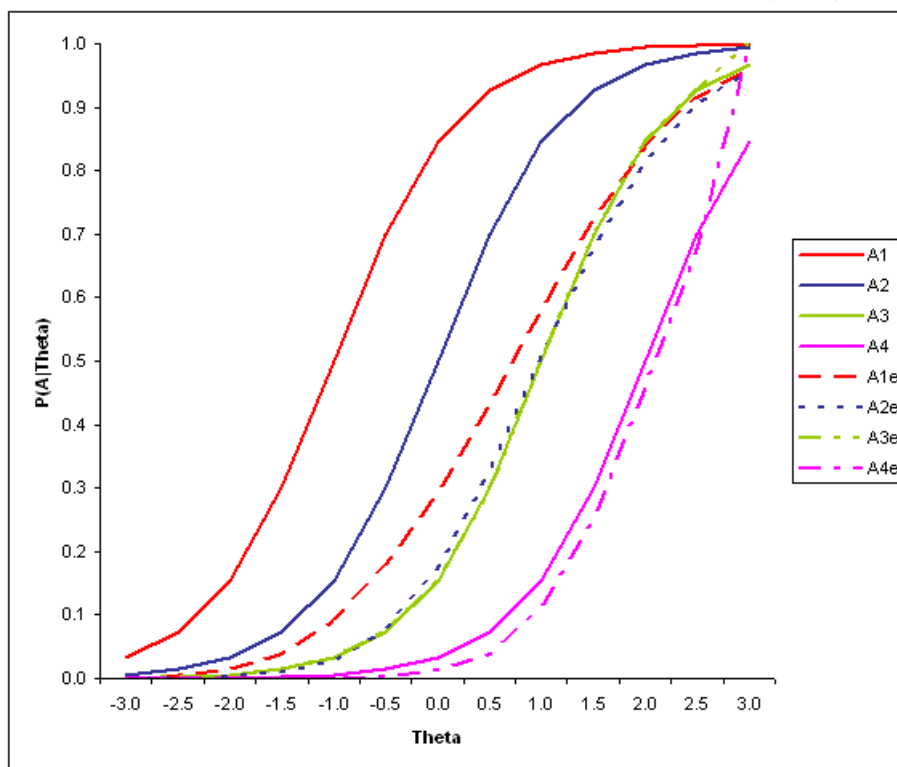


Las CPAs de la condición SFE no se han visto afectadas (ver Gráfica 4.5), con valores RMSE muy bajos: 0.000 (A_1), 0.026 (A_2), 0.014 (A_3) y 0.004 (A_4). En la Gráfica 4.6 se puede observar que el atributo más afectado de la condición DFE ha sido A_1 pues el método presenta una tendencia a infraestimar sus probabilidades de dominio. Los valores RMSE para dicha condición han sido 0.250 (A_1), 0.081 (A_2), 0.000 (A_3) y 0.008 (A_4).

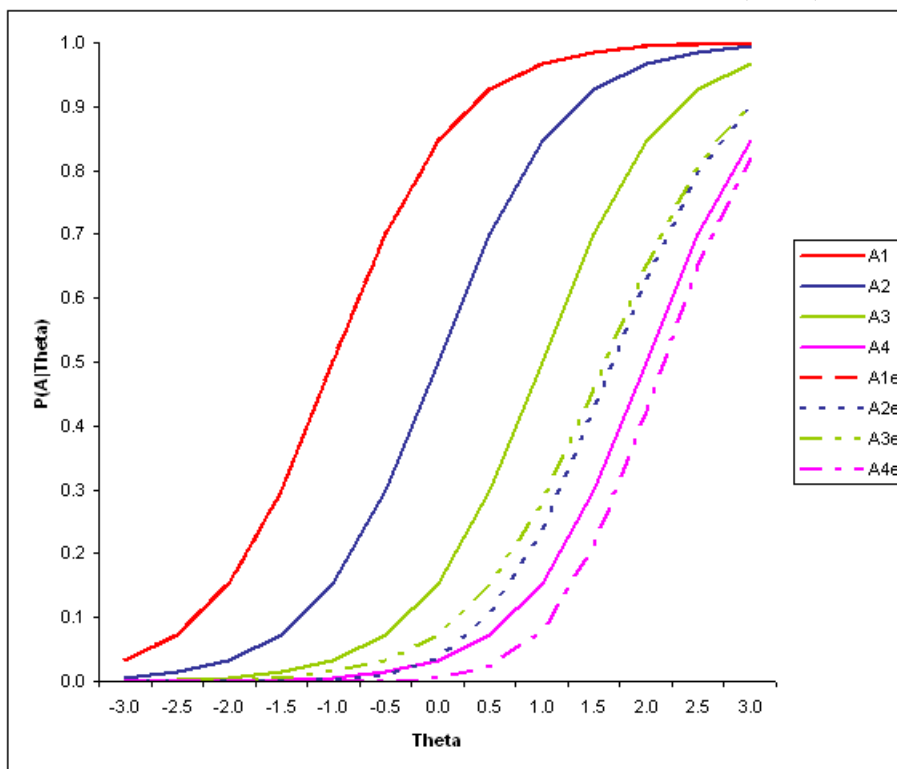
En las Gráficas 4.7 y 4.8 se presentan las CPAs correspondientes a las condiciones de *exclusión* en atributos difíciles para las matrices esparcida y densa respectivamente (condiciones SDE y DDE). En dichas gráficas se puede observar que los atributos más afectados en la condición SDE han sido A_1 y A_2 , aunque no han recibido ningún cambio. El LSDM tiende a subestimar las probabilidades de dominio de dichos atributos. Los valores RMSE de esta condición han sido: 0.322 (A_1), 0.202 (A_2), 0.009 (A_3) y 0.049 (A_4).

Finalmente, en la condición DDE (Gráfica 4.8) se puede observar que los atributos más afectados son A_2 y A_3 , cuyas probabilidades de dominio han sido infraestimadas por el método LSDM. Los valores RMSE para dicha condición han sido: 0.000 (A_1), 0.337 (A_2), 0.123 (A_3) y 0.045 (A_4). En general, los resultados indican que las condiciones con mayor afectación de las curvas han sido las que tienen errores de especificación en los atributos difíciles (SDE, DDE y SDI), siendo especialmente notorio el cambio en las probabilidades de los atributos de la condición de *inclusión* (SDI).

Gráfica 4.7: CPAs condición esparsa, difícil, excluir (SDE)



Gráfica 4.8: CPAs condición densa, difícil, excluir (DDE)



4.3.1.2. Errores de adición y fusión de atributos

Análisis de los efectos en la recuperación de las CCIs

En la Tabla 4.13 se presentan los estadísticos de recuperación de las CCIs obtenidos tras la aplicación del LSDM en las condiciones en las que se han fusionado atributos. En dicha tabla se ha señalado con un asterisco los ítems mal recuperados según los puntos de corte expuestos en la tabla 3.7 del Capítulo 3. En los resultados expuestos no se presentan los valores MAD de las condiciones de adición de atributos (SA y DA) pues han sido iguales a cero para todos los ítems en dichas condiciones. Lo anterior indica que adicionar un atributo no relevante para la solución de los ítems no produce ningún efecto en la recuperación de las CCIs con el método LSDM.

Tabla 4.13: Estadísticos MAD para evaluar la recuperación de las CCIs

	SFAF	SDAF	SFOF	SDOF	DFAF	DDAF	DFOF	DDOF
i1	0.116*	0.074*	0.001	0.262*	0.001	0.023	0.002	0.008
i2	0.035	0.049	0.001	0.155*	0.003	0.019	0.002	0.009
i3	0.002	0.111*	0.010	0.052	0.000	0.005	0.001	0.057*
i4	0.000	0.034	0.003	0.047	0.000	0.120*	0.002	0.062*
i5	0.069	0.065*	0.000	0.203*	0.002	0.024*	0.000	0.014
i6	0.014	0.101*	0.002	0.055				
i7	0.004	0.036	0.000	0.190*				
i8	0.009	0.083*	0.023	0.033				
i9	0.003	0.038	0.007	0.152*				
i10	0.000	0.066	0.002	0.000				

* ítems mal recuperados según punto de corte.

Los resultados de la fusión de atributos, presentados en la Tabla 4.13, muestran que al unir atributos difíciles se produce un empeoramiento de la recuperación de las CCIs con respecto a la condición base (**Q** verdadera); esto se refleja en que los valores MAD de aproximadamente la mitad de los ítems de las condiciones SDAF, SDOF, DDAF y DDOF han superado el punto de corte correspondiente, evidenciando una mala recuperación. Sin embargo no ocurre lo mismo cuando se mezclan los atributos fáciles (condiciones SFAF, SFOF, DFAF y DFOF), pues no se observa un aumento notable en los estadísticos MAD, indicando que este tipo de error no produce efecto en la recuperación de las CCIs. No se aprecian diferencias en cuanto a la recuperación de las CCIs según la forma de fusionar los atributos (*AND* u *OR*).

Análisis de los valores residuales LSD

En la Tabla 4.14 se presentan los valores LSD para cada condición. En dicha tabla se puede observar que los valores LSD son muy altos en las condiciones que involucran la fusión de atributos difíciles (SDAF, SDOF, DDAF, DDOF), tanto si la matriz \mathbf{Q} es densa como si es esparcida. Las condiciones que registran un menor aumento de los valores residuales son las condiciones en que se fusionan atributos fáciles de forma inclusiva. Nuevamente, los valores residuales son más altos en los niveles bajos de habilidad y disminuyen progresivamente cuando aumenta la habilidad. En las condiciones de adición de un atributo irrelevante (DA y SA) tampoco se aprecia un efecto sobre los valores LSD pues su estimación ha sido 0 en todos los niveles de habilidad.

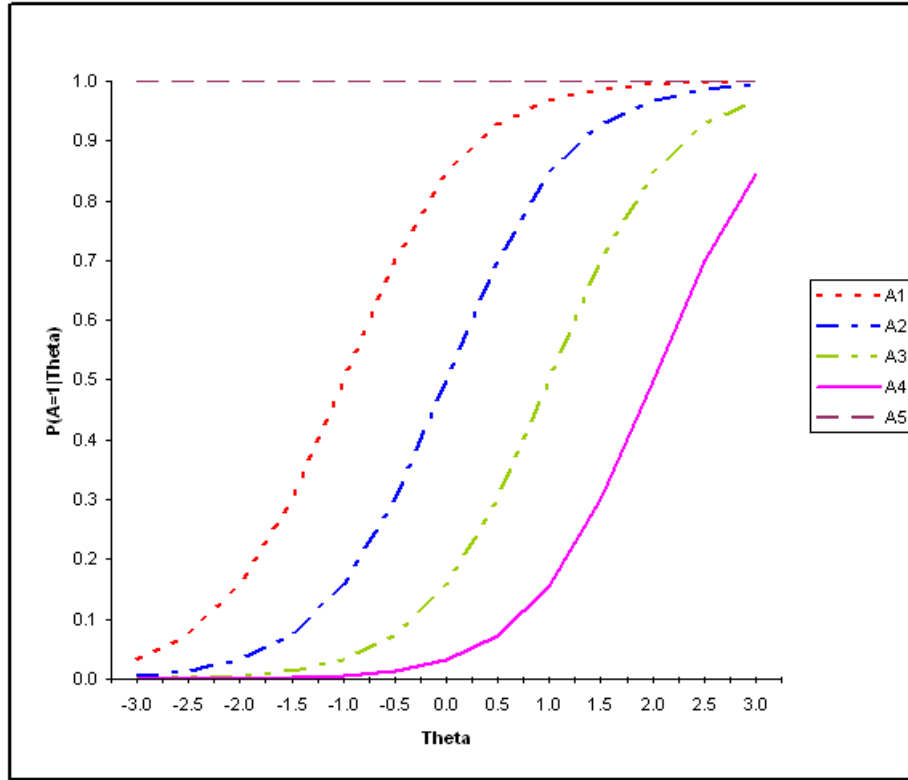
Tabla 4.14: Valores residuales LSD en cada nivel fijo de habilidad

	SFAF	SDAF	SFOF	SDOF	DFOF	DDAF	DFAF	DDOF
-3	0.436	0.721	0.852	1.497	0.736	1.304	0.410	0.646
-2.5	0.370	0.648	0.697	1.334	0.608	1.183	0.356	0.585
-2	0.306	0.576	0.549	1.172	0.486	1.065	0.303	0.525
-1.5	0.245	0.505	0.411	1.011	0.371	0.948	0.250	0.466
-1	0.184	0.436	0.288	0.852	0.267	0.829	0.194	0.410
-0.5	0.125	0.370	0.184	0.697	0.175	0.703	0.134	0.356
0	0.075	0.306	0.106	0.549	0.102	0.572	0.081	0.303
0.5	0.039	0.245	0.054	0.411	0.053	0.440	0.043	0.250
1.0	0.018	0.185	0.025	0.288	0.025	0.314	0.020	0.194
1.5	0.008	0.126	0.011	0.184	0.011	0.204	0.009	0.134
2	0.003	0.075	0.005	0.106	0.005	0.117	0.004	0.081
2.5	0.001	0.039	0.002	0.054	0.002	0.060	0.002	0.043
3	0.000	0.018	0.001	0.025	0.001	0.028	0.001	0.020

Análisis de los efectos sobre las CPAs

La Gráfica 4.9 presenta las CPAs estimadas por el LSDM tras adicionar un atributo irrelevante en la matriz \mathbf{Q} .

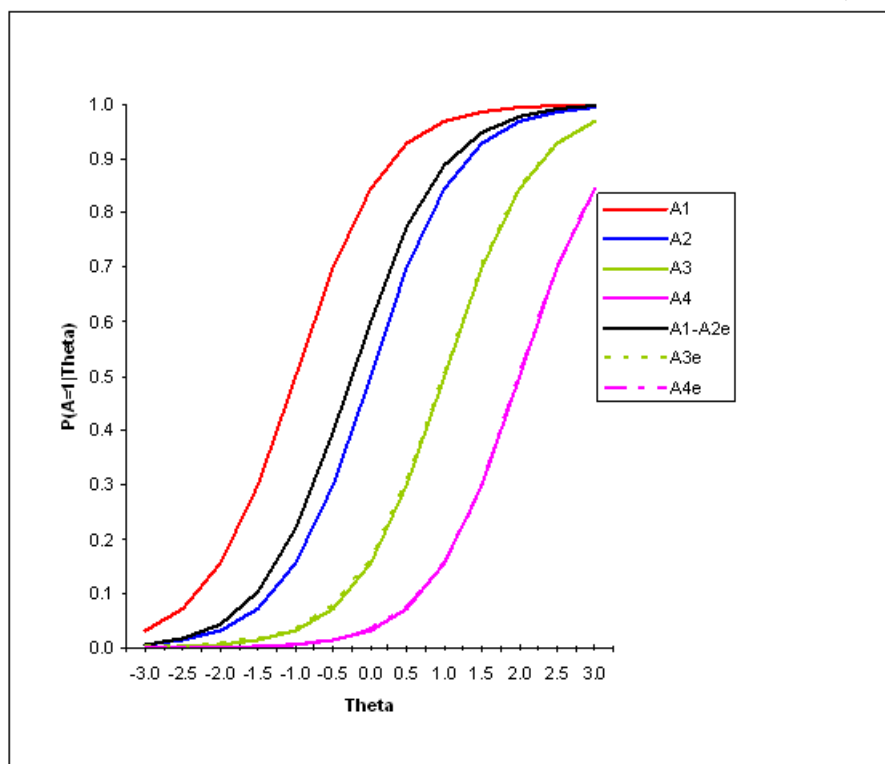
Gráfica 4.9: CPAs de las condiciones de adición (SA y DA)



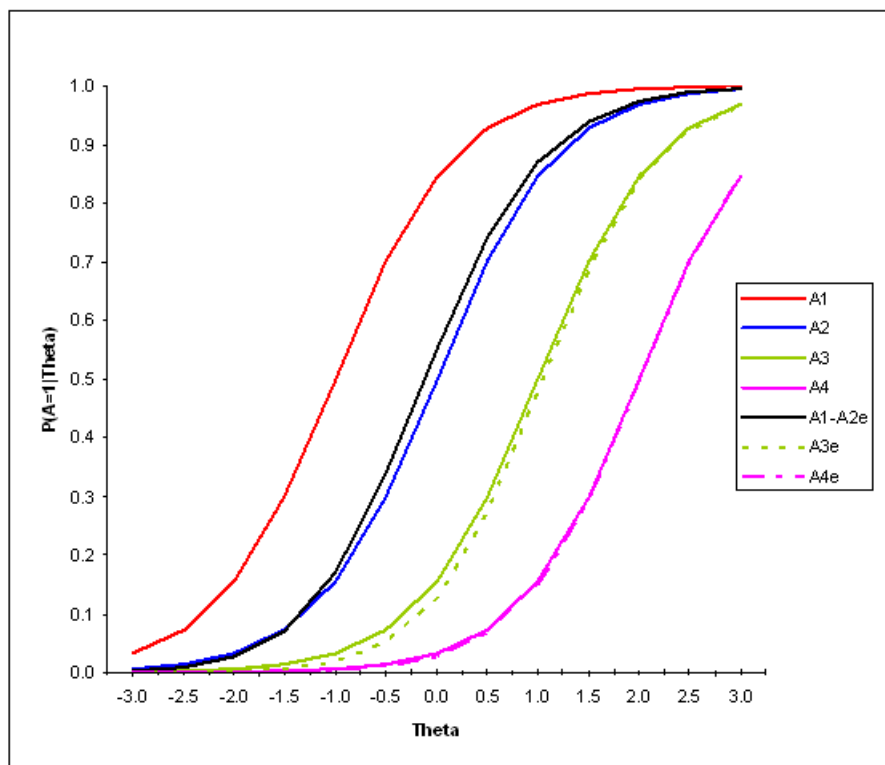
En la Gráfica 4.9 se puede observar que las probabilidades del atributo adicionado son 1 en todos los niveles de habilidad, tanto para la matriz densa (DA) como para la esparcida (SA), y las CPAs de los atributos restantes no se ven alteradas, lo anterior puede ser explicado porque la estimación del vector \mathbf{X} mediante el método LSDM es 0 para dicha columna y al ser multiplicado por \mathbf{Q} no altera las restantes probabilidades (el vector nulo actúa como módulo en la multiplicación de las matrices \mathbf{QX}) por ello es la solución que minimiza la norma del vector $\mathbf{QX}-\mathbf{L}$. Como la CPA es el valor exponencial del vector \mathbf{X} las probabilidades de dominio son 1, sin embargo, dicha probabilidad es consecuencia de la solución algebraica sin que se pueda interpretar como una estimación real del dominio del atributo. Este resultado tiene importantes implicaciones para los estudios con datos reales ya que la presencia de este comportamiento en algún atributo indica que el componente en cuestión no aporta información a la habilidad para resolver los ítems.

Las Gráficas 4.10 y 4.11 exhiben los resultados de las condiciones en que se fusionan atributos fáciles mediante la regla *AND*, para las matrices \mathbf{Q} esparcida y densa respectivamente.

Gráfica 4.10: CPAs de la condición esparcida, fácil, AND, fusionar (SFAF)



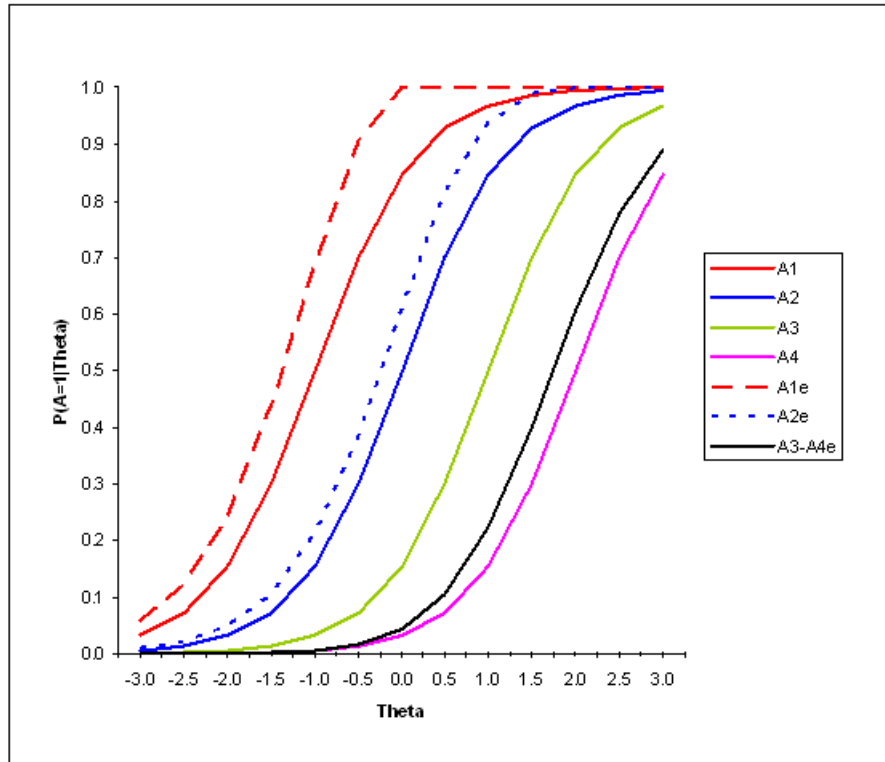
Gráfica 4.11: CPAs de la condición densa, fácil, AND, fusionar (DFAF)



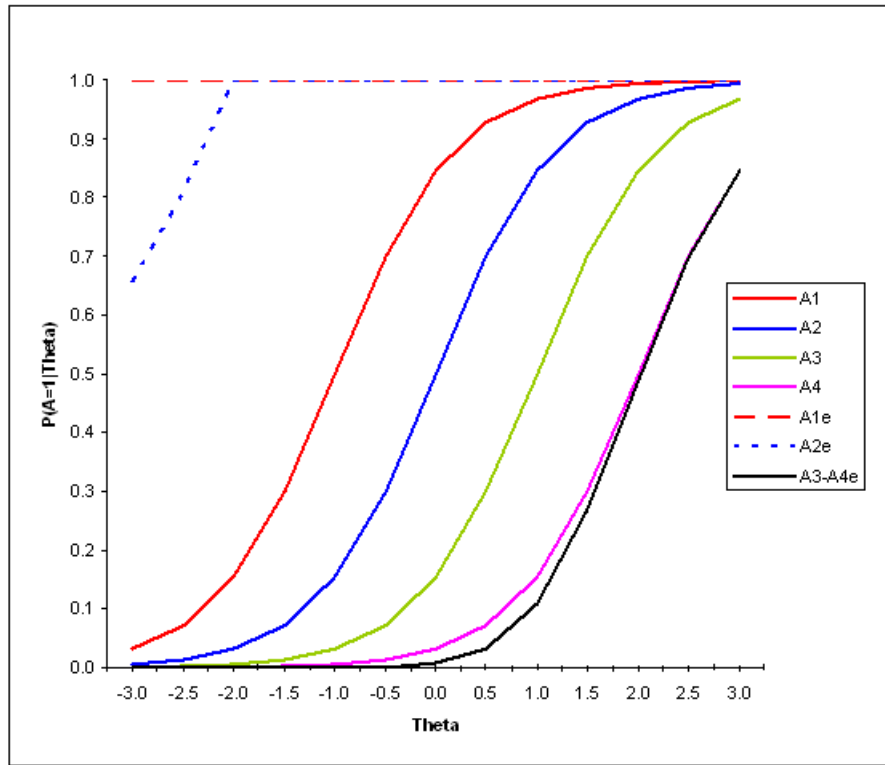
En las gráficas se puede observar que la estimación del atributo fusionado se aproxima a la curva del atributo A_2 y que las probabilidades de los atributos no manipulados no se afectan, los valores RMSE confirman lo anteriormente expuesto: 0.158 (A_1), 0.050 (A_2), 0.004 (A_3) y 0.001 (A_4). Es importante notar que se presentan cuatro valores RMSE ya que se comparan las curvas originales de A_1 y A_2 con el atributo fusionado $A_1 - A_2$. Los resultados de la manipulación de la matriz \mathbf{Q} densa (DFAF) son muy similares a los encontrados en la esparcida con valores de probabilidad de A_3 y A_4 poco alterados y con el atributo fusionado más cercano a A_2 que a A_1 , los valores RMSE de dicha condición son: 0.187 (A_1), 0.023 (A_2), 0.016 (A_3) y 0.004 (A_4).

En las Gráficas 4.12 y 4.13 se presentan las estimaciones de las CPAs en las condiciones de fusión de atributos difíciles, mediante la regla *AND*, para las matrices esparcida y densa respectivamente.

Gráfica 4.12: CPAs de la condición esparcida, difícil, AND, fusionar (SDAF)



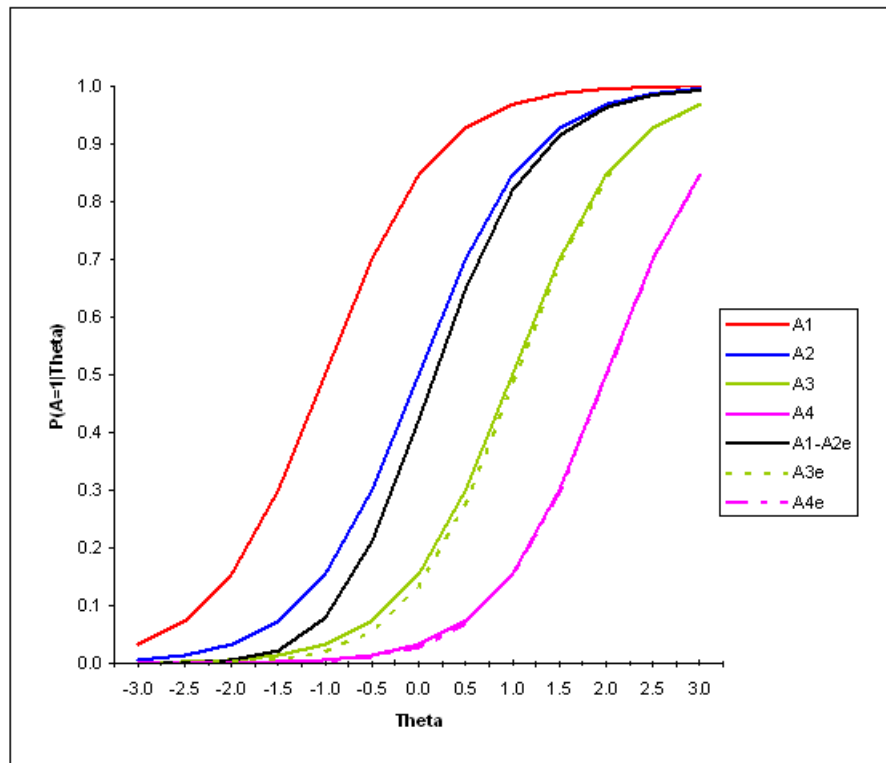
Gráfica 4.13: CPAs de la condición densa, difícil, AND, fusionar (DDAF)



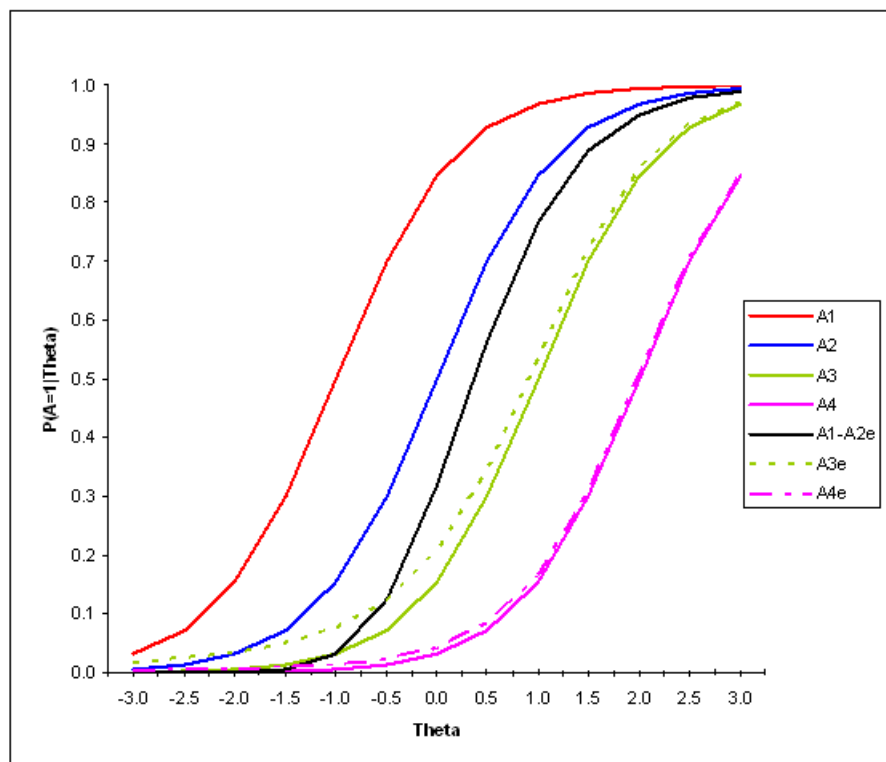
En las Gráficas 4.12 y 4.13 se puede observar que las condiciones en las que se fusionan los atributos difíciles producen una fuerte alteración en las CPAs, especialmente en la matriz \mathbf{Q} densa. Además no solo se encuentran afectadas las probabilidades de los atributos fusionados, sino también las de los atributos en los que no se han introducido especificaciones erróneas, en los que se sobreestima las probabilidades de dominio. Los valores RMSE han sido: $A_1 = 0.101$, $A_2 = 0.631$, $A_3 = 0.154$ y $A_4 = 0.052$ en la condición SDAF y 0.509 (A_1), 0.584 (A_2), 0.221 (A_3) y 0.021 (A_4) en la condición DDAF.

Las Gráficas 4.14 y 4.15 muestran los resultados de las condiciones en que se fusionan atributos fáciles mediante la regla *OR* para las matrices \mathbf{Q} esparcida y densa respectivamente.

Gráfica 4.14: CPAs de la condición esparcida, fácil, OR, fusionar (SFOF)



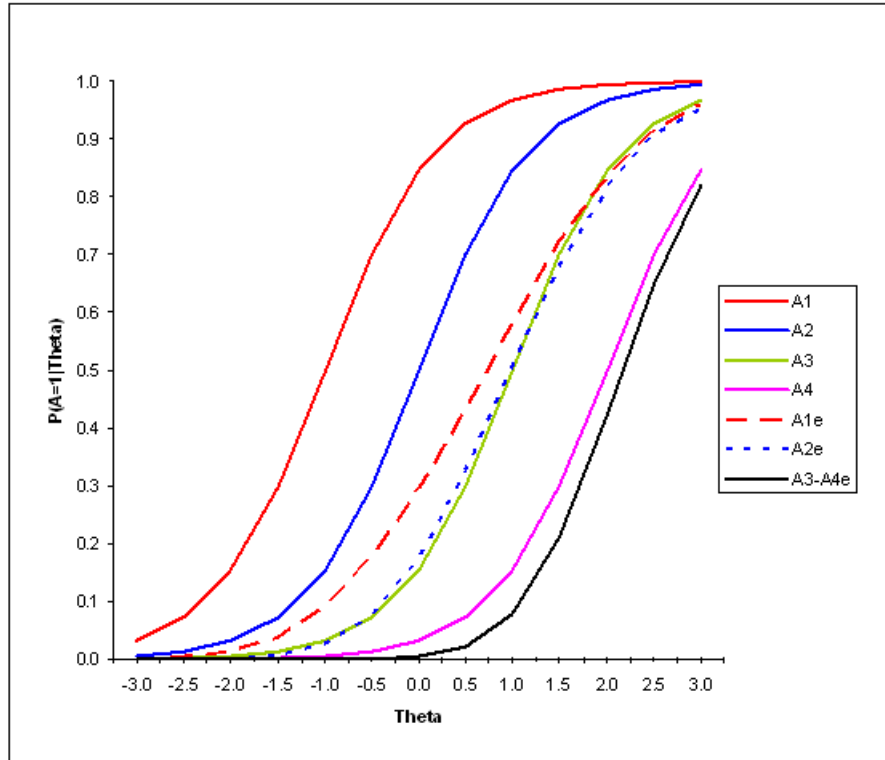
Gráfica 4.15: CPAs de la condición densa, fácil, OR, fusionar (DFOF)



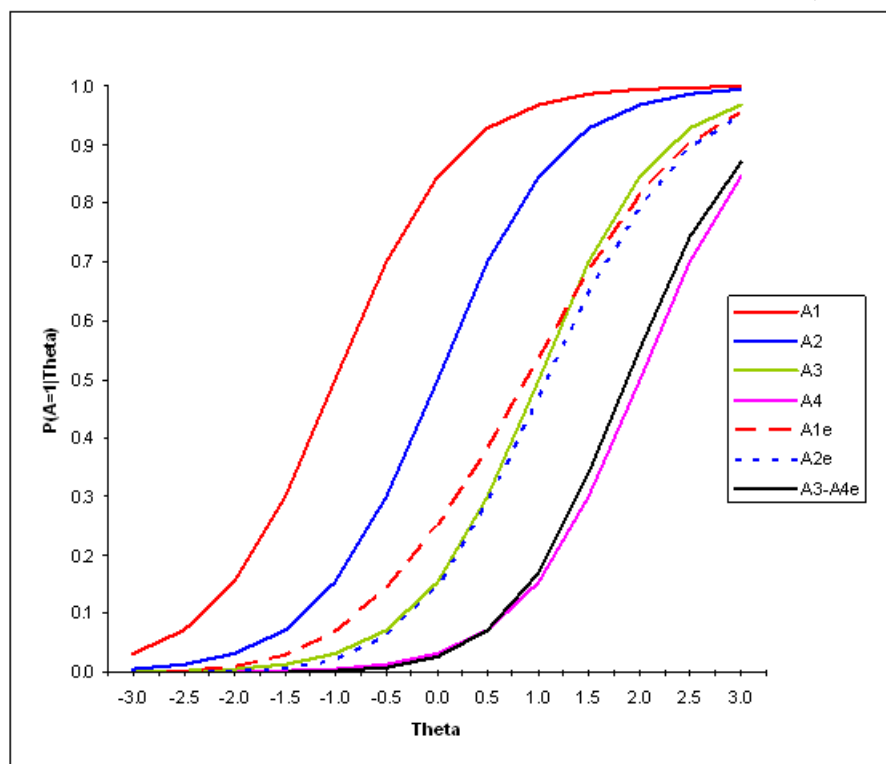
En las Gráficas 4.14 y 4.15 se puede apreciar una tendencia muy similar a la de la regla *AND* pues la fusión de atributos fáciles no afecta la estimación de las CPAs de A_3 y A_4 . Además, el atributo fusionado se asemeja más a A_2 que a A_1 . Los valores RMSE han sido 0.252 (A_1), 0.046 (A_2), 0.014 (A_3) y 0.040 (A_4) para condición SFOF y 0.297 (A_1), 0.093 (A_2), 0.033 (A_3) y 0.008 (A_4) para la condición DFOF. Se observa también en los resultados que las condiciones en las que se usa la regla *OR* producen un mayor efecto en las CPAs que las que aplican la regla *AND* para ejecutar la fusión.

En las Gráficas 4.16 y 4.17 se exhiben las CPAs de las condiciones en que se fusionan atributos difíciles mediante la regla *OR* para las matrices \mathbf{Q} esparcida y densa respectivamente. Como se puede observar en las gráficas la fusión de los atributos difíciles (A_{3-4}) produce importantes efectos en los atributos fáciles (A_1 y A_2), con una notable infrastimulación de las probabilidades de dominio para estos últimos. Lo anterior se confirma con los valores RMSE que son, $A_1 = 0.331$, $A_2 = 0.201$, $A_3 = 0.250$ y $A_4 = 0.045$ para la condición SDOF y 0.351 (A_1), 0.221 (A_2), 0.185 (A_3) y 0.023 (A_4) para la condición DDOF.

Gráfica 4.16: CPAs de la condición esparcida, difícil, OR, fusionar (SDOF)



Gráfica 4.17: CPAs de la condición densa, difícil, OR, fusionar (DDOF)



En general se aprecia un mayor efecto de los errores de especificación en las condiciones que involucran cambios en los atributos difíciles, existiendo una tendencia a la sobrestimación de las probabilidades cuando utiliza la regla *AND* para fusionar la pareja de atributos y a la infraestimación cuando se aplica la regla *OR*.

4.3.1.3. Errores de dependencia e incompatibilidad entre atributos

Análisis de los efectos en la recuperación de las CCIs

En la Tabla 4.15 se presentan los valores MAD obtenidos tras la aplicación del LSDM en las condiciones en las que se han introducido errores de dependencia secuencial e incompatibilidades entre atributos, se ha señalado con un asterisco los ítems mal recuperados ya que exceden los puntos de corte establecidos. Las siglas para cada condición son: SFR1 (esparcida, fácil, dependencia simple), SDR1 (esparcida, difícil, dependencia simple), DFR1 (densa, fácil, dependencia simple), DDR1 (densa, difícil, dependencia simple), SFI (esparcida, fácil, incompatibilidad), SDI (esparcida, difícil, incompatibilidad), DFI (densa, fácil, incompatibilidad), DDI (densa, difícil, incompatibilidad), SR2 (esparcida, dependencia

compleja) y DR2 (densa, dependencia compleja).

Tabla 4.15: Estadísticos MAD para evaluar la recuperación de las CCI's

	SFR1	SDR1	SFI	SDI	SR2	DFR1	DDR1	DFI	DDI	DR2
i1	0.006	0.054	0.081*	0.077*	0.000	0.001	0.006	0.000	0.023	0.017
i2	0.012	0.026	0.006	0.056	0.011	0.000	0.002	0.002	0.019	0.012
i3	0.001	0.006	0.003	0.077*	0.018	0.000	0.020	0.000	0.000	0.000
i4	0.000	0.011	0.001	0.003	0.005	0.000	0.000	0.000	0.120*	0.001
i5	0.022	0.039	0.105*	0.074*	0.022*	0.001	0.007	0.002	0.024*	0.001
i6	0.000	0.002	0.007	0.067*	0.008					
i7	0.000	0.009	0.002	0.001	0.002					
i8	0.001	0.003	0.001	0.049*	0.016					
i9	0.000	0.007	0.000	0.004	0.005					
i10	0.000	0.021	0.001	0.101*	0.001					

En la Tabla 4.15 se puede observar que la condición que produce mayor efecto en la recuperación de los ítems es la incompatibilidad entre atributos difíciles en ambos tipos de **Q** (densa y esparcida). En el caso de matriz **Q** esparcida también ocurre algún deterioro en la recuperación de las CCI's cuando se introducen incompatibilidades entre atributos fáciles. Contrario a lo que se esperaba, las relaciones de dependencia secuencial tanto simples como complejas no han tenido un efecto relevante en la recuperación de las CCI's según el estadístico MAD.

Análisis de los valores residuales LSD

En la Tabla 4.16 se presentan los valores LSD para cada condición. Se puede notar que dichos valores son más altos en las condiciones que involucran dependencias e incompatibilidades en las matrices densas. Los valores residuales más altos se han presentado en las condiciones de incompatibilidad entre atributos difíciles (SDI y DDI), aunque las incompatibilidades entre atributos fáciles también incrementan notablemente los valores LSD. Las relaciones de dependencia secuencial, tanto simple (SDR1, SFR1, DFR1 y DDR1) como compleja (SR2 y DR2), tienen un efecto menor en el aumento de los residuales pero, aún así, los LSD son notablemente grandes. Al igual que en los escenarios previos, los valores LSD tienden a disminuir a medida que aumenta el nivel de habilidad.

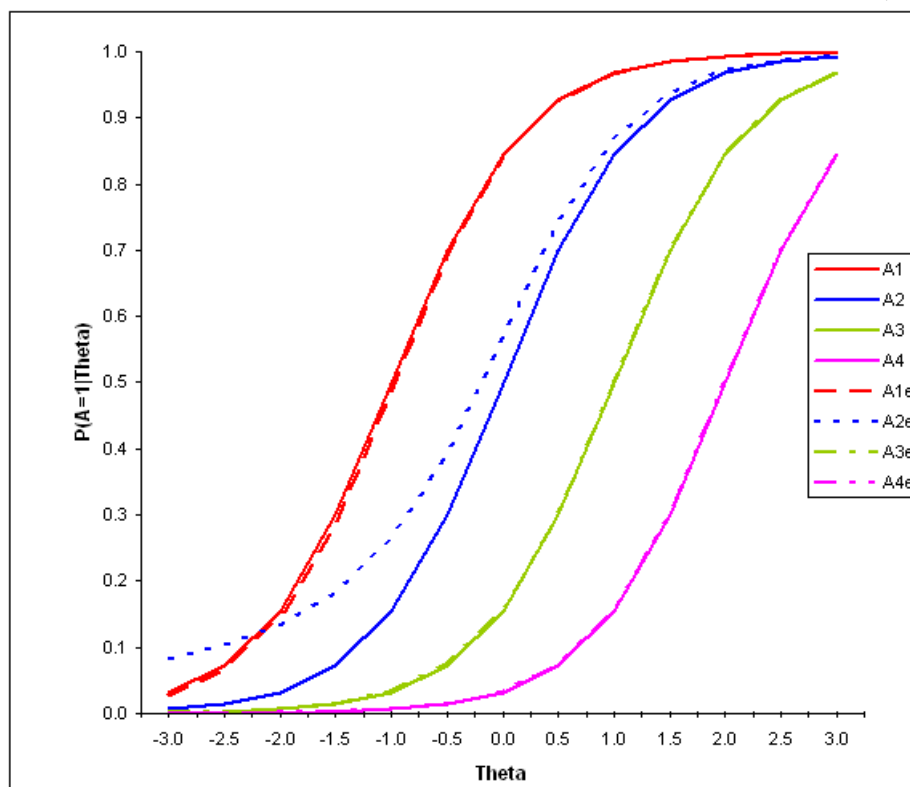
Tabla 4.16: Valores residuales LSD en cada nivel fijo de habilidad

	SFR1	SDR1	SFI	SDI	SR2	DFR1	DDR1	DFI	DDI	DR2
-3	0.291	0.576	0.432	0.720	0.579	0.485	0.964	0.722	1.269	0.639
-2.5	0.222	0.504	0.361	0.648	0.473	0.371	0.843	0.603	1.009	0.583
-2	0.158	0.432	0.290	0.576	0.372	0.264	0.730	0.485	0.993	0.492
-1.5	0.102	0.361	0.222	0.504	0.280	0.170	0.621	0.371	0.890	0.390
-1	0.058	0.290	0.158	0.432	0.198	0.098	0.513	0.264	0.783	0.235
-0.5	0.030	0.222	0.102	0.361	0.128	0.050	0.402	0.170	0.671	0.158
0	0.014	0.158	0.058	0.290	0.074	0.023	0.292	0.98	0.554	0.094
0.5	0.006	0.102	0.030	0.223	0.038	0.010	0.191	0.050	0.443	0.048
1.0	0.003	0.058	0.014	0.159	0.018	0.004	0.111	0.023	0.312	0.038
1.5	0.001	0.030	0.006	0.102	0.008	0.002	0.057	0.010	0.204	0.018
2	0.000	0.014	0.002	0.059	0.003	0.001	0.027	0.004	0.117	0.013
2.5	0.000	0.006	0.001	0.030	0.001	0.000	0.001	0.002	0.060	0.001
3	0.000	0.002	0.000	0.014	0.000	0.000	0.000	0.001	0.028	0.001

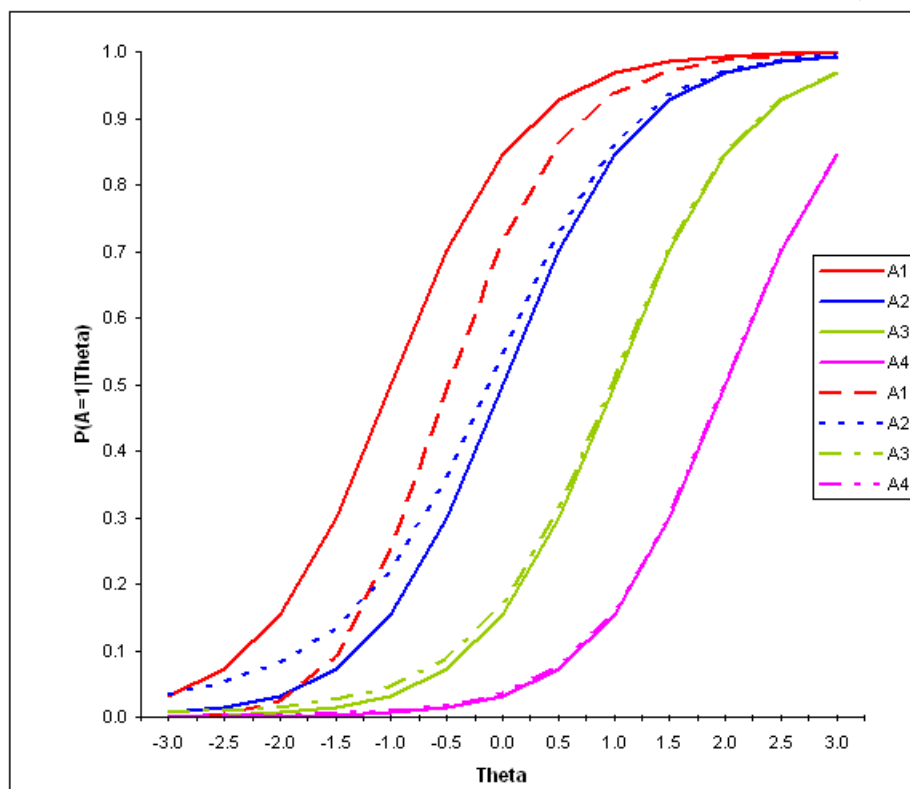
Análisis de los efectos sobre las CPAs

En las Gráficas 4.18 y 4.19 se presentan las CPAs estimadas con el LSDM para las condiciones en las que se han introducido errores de dependencia simple entre atributos fáciles. En dichas gráficas se puede apreciar que este tipo de error de especificación produce pocas alteraciones en las CPAs, y que dichos efectos se producen en los atributos A_1 y A_2 , que son los atributos que han recibido error de especificación y no en A_3 y A_4 cuya estimación ha sido muy precisa, con valores RMSE de 0.009 (A_1), 0.070 (A_2), 0.001 (A_3) y 0 (A_4) para SFR1 y 0.124 (A_1), 0.038 (A_2), 0.009 (A_3) y 0.002 (A_4) para DFR1.

Gráfica 4.18: CPAs de la condición esparcida, fácil, dependencia simple (SFR1)

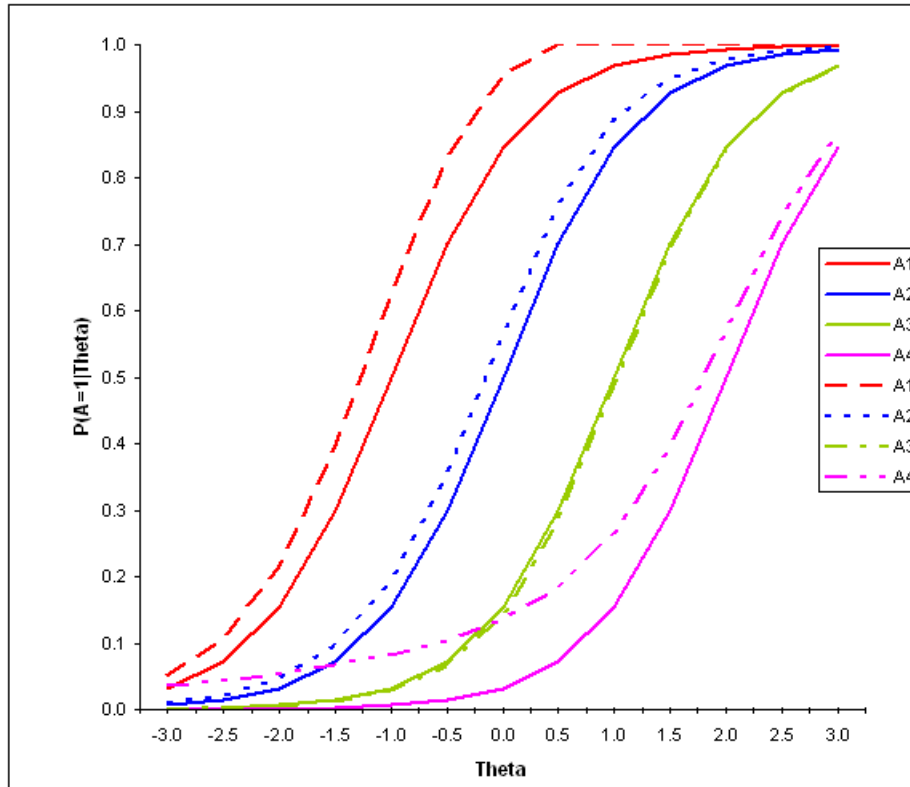


Gráfica 4.19: CPAs de la condición densa, fácil, dependencia simple (DFR1)

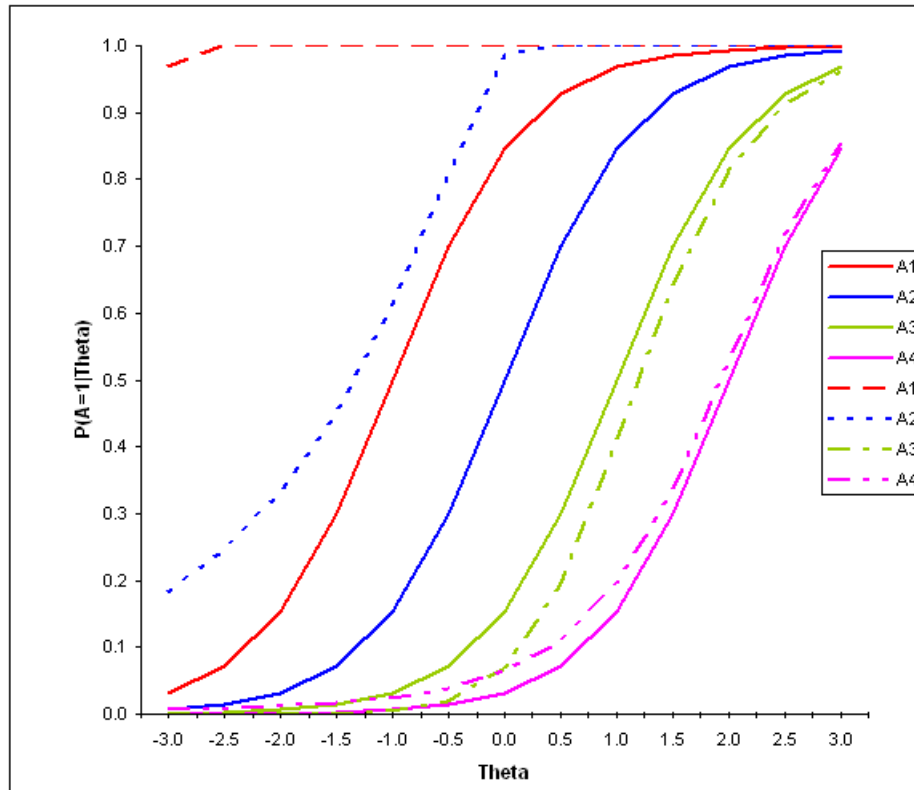


Las Gráficas 4.20 y 4.21 exhiben las CPAs de las condiciones en las que se han introducido errores de dependencia simple entre atributos de alta dificultad relativa.

Gráfica 4.20: CPAs de la condición esparcida, difícil, dependencia simple (SDR1)



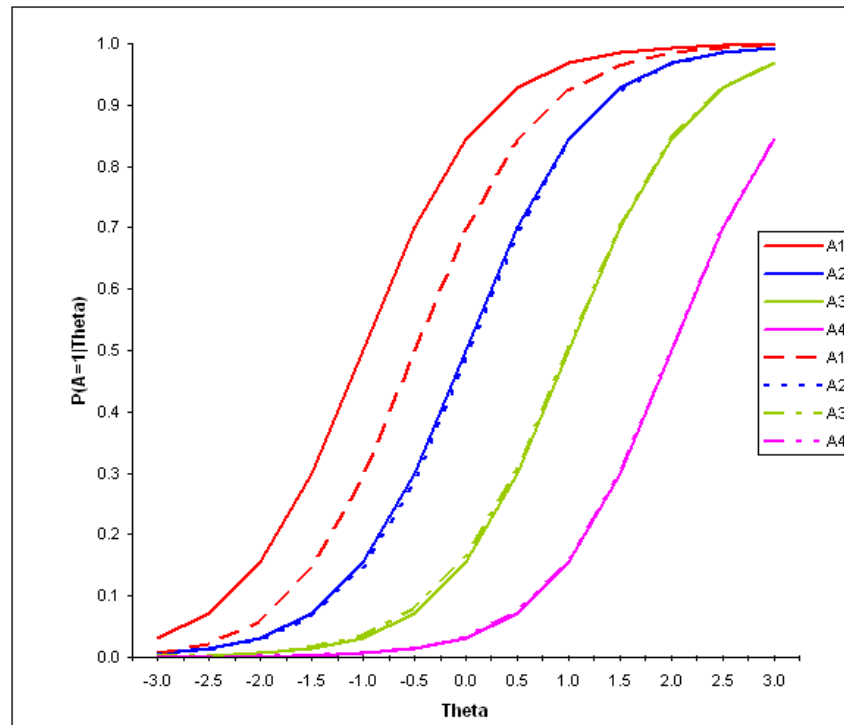
Gráfica 4.21: CPAs de la condición densa, difícil, dependencia simple (DDR1)



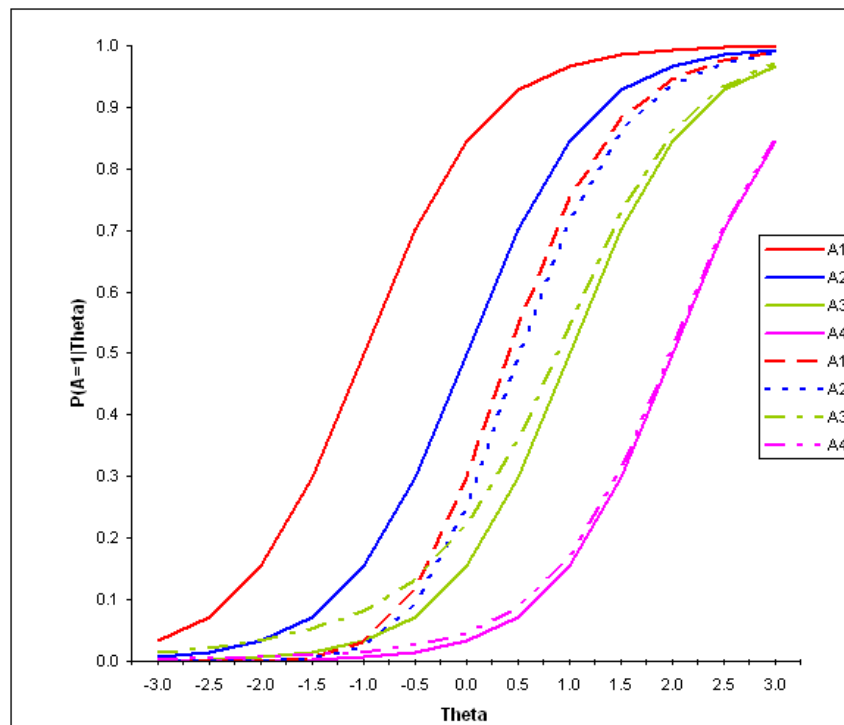
En las Gráficas 4.20 y 4.21 se puede observar que la introducción de este tipo de error afecta principalmente la estimación de A_1 y A_2 en la matriz densa con una notoria sobrestimación de las probabilidades de dominio del atributo A_1 , esto se ve confirmado con valores RMSD de 0.505 (A_1), 0.296 (A_2), 0.052 (A_3) y 0.024 (A_4) para la condición DDR1. La matriz esparcida (SDR1) no se ve tan afectada por este tipo de error de especificación y presenta valores RMSD de 0.070 (A_1), 0.034 (A_2), 0.008 (A_3) y 0.071 (A_4).

Las Gráficas 4.22 y 4.23 presentan las CPAs estimadas en las condiciones de incompatibilidad entre atributos fáciles para la matriz \mathbf{Q} esparcida y densa respectivamente. En estas figuras se puede observar que la incompatibilidad entre atributos fáciles tiene efecto en la estimación de los atributos afectados (A_1 , A_2), infraestimando sus probabilidades de dominio. Al igual que en las condiciones precedentes la estimación de los atributos A_3 y A_4 no se ha visto afectada. Se nota también un efecto mayor en la matriz densa que en la esparcida. Los valores RMSE han sido: $A_1 = 0.110$, $A_2 = 0.008$, $A_3 = 0.004$ y $A_4 = 0.001$ para la condición SFI y 0.305 (A_1), 0.123 (A_2), 0.038 (A_3) y 0.009 (A_4) para la condición DFI.

Gráfica 4.22: CPAs de la condición esparcida, fácil, incompatibilidad (SFI)



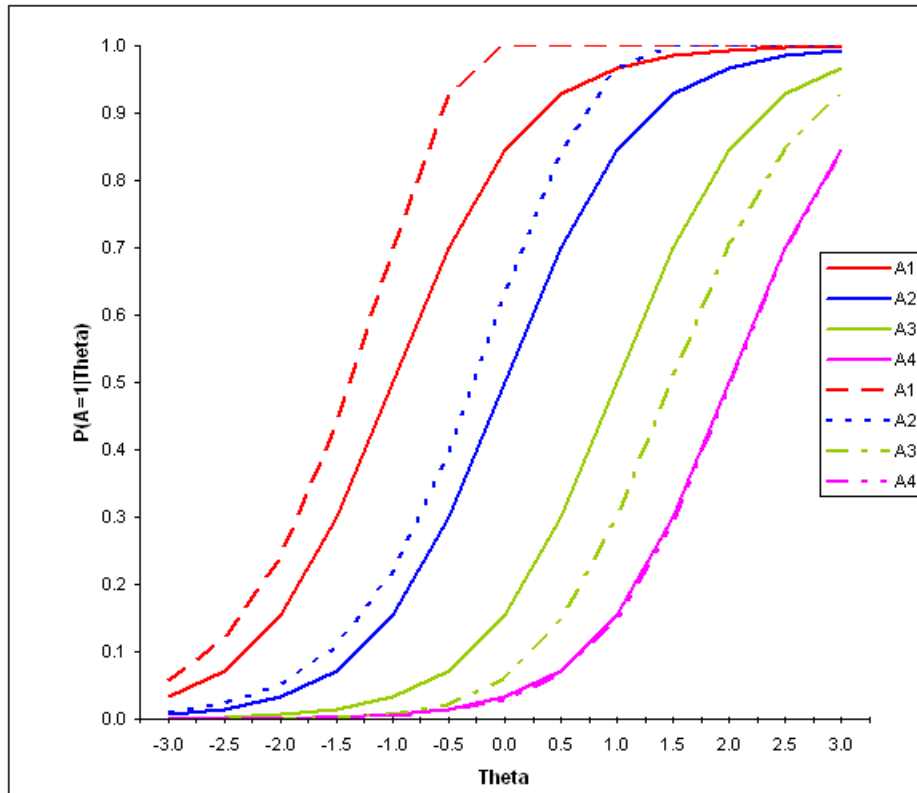
Gráfica 4.23: CPAs de la condición densa, fácil, incompatibilidad (DFI)



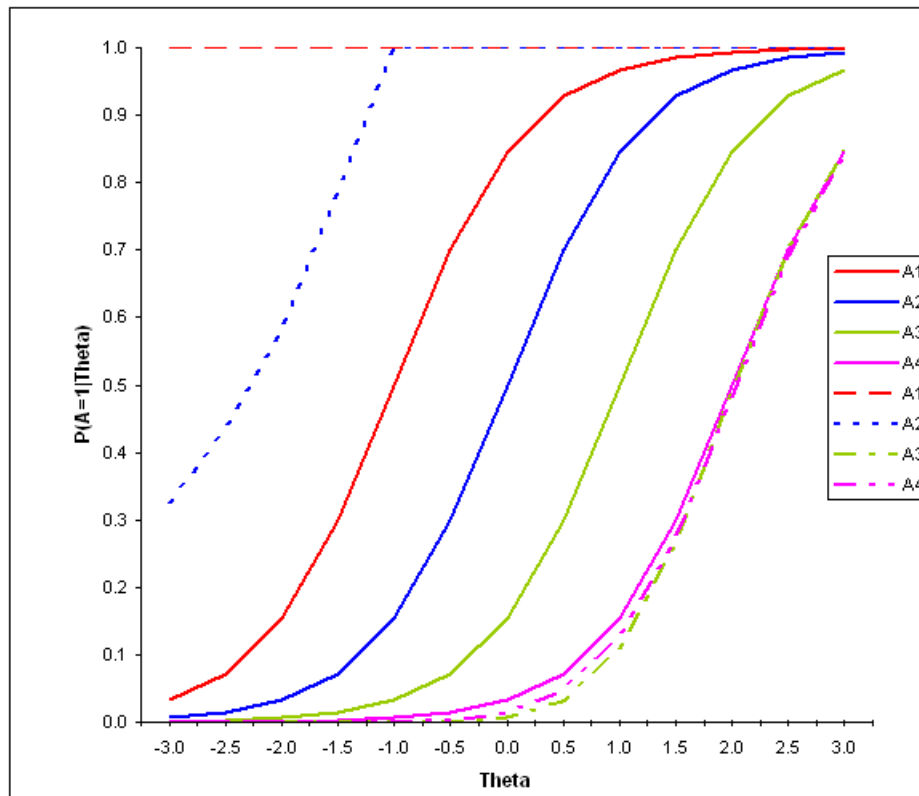
En las Gráficas 4.24 y 4.25 se exhiben las CPAs estimadas en las condiciones de incompatibilidad entre atributos difíciles para la matriz \mathbf{Q} esparcida y densa respectivamente.

En dichas gráficas se puede apreciar que las condiciones de incompatibilidad entre atributos difíciles son las que mayores cambios producen en la forma y orden de las CPAs, afectando principalmente el atributo A_1 . Esta tendencia se confirma en los valores RMSE: $A_1 = 0.107$, $A_2 = 0.074$, $A_3 = 0.104$ y $A_4 = 0.005$ de la condición SDI. Cuando la matriz es densa, los atributos con mayor afectación son A_1 y A_2 ya que sus probabilidades tienden a ser sobreestimadas por el LSDM, llegando incluso a ser 1 en todos los niveles de habilidad; dicho resultado se ve confirmado con valores RMSE de 0.509 (A_1), 0.453 (A_2), 0.222 (A_3) y 0.016 (A_4) para la condición DDI.

Gráfica 4.24: CPAs de la condición esparcida, difícil, incompatibilidad (SDI)

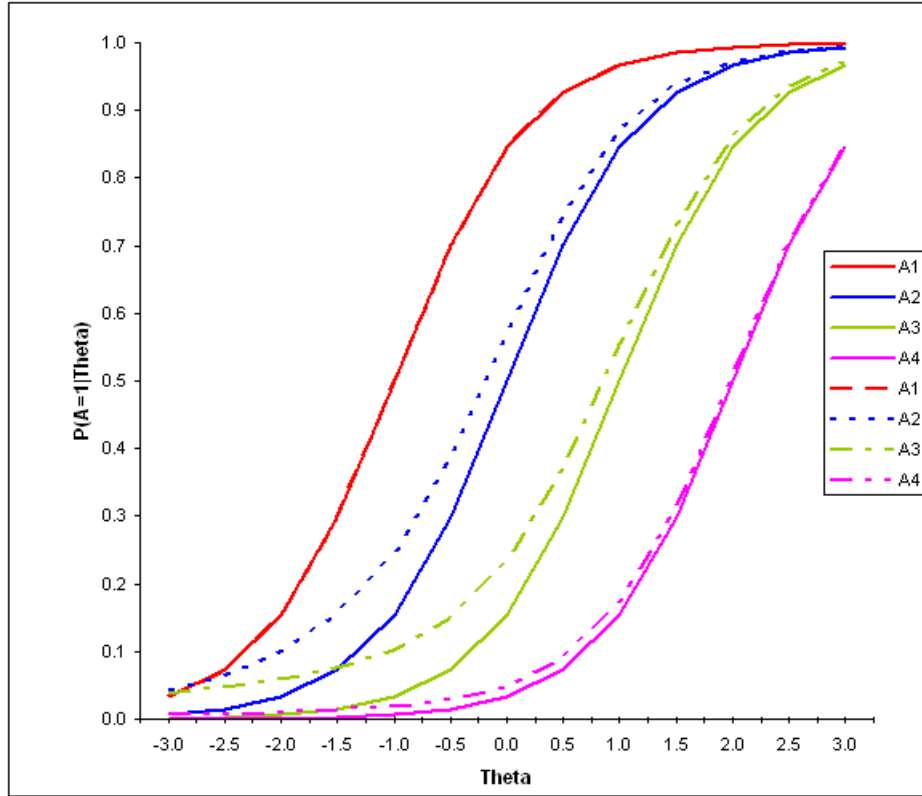


Gráfica 4.25: CPAs de la condición densa, difícil, incompatibilidad (DDI)



Finalmente, en la Gráfica 4.26 se puede observar que en la condición de dependencia condicional compleja (SR2) no existe un efecto notable en la estimación de las CPAs, esto se confirma con valores RMSE de 0.006 (A_1), 0.054 (A_2), 0.052 (A_3) y 0.012 (A_4).

Gráfica 4.26: CPAs de la condición esparcida, dependencia compleja (SR2)



4.3.2. Errores de heterogeneidad y falta de monotonicidad

En este apartado se presentan los resultados de la heterogeneidad y de la no monotonicidad de los atributos.

4.3.2.1. Análisis de los efectos de la heterogeneidad

La Tabla 4.17 presenta los valores MAD para evaluar la calidad de la recuperación de los ítems, la matriz \mathbf{Q} y los valores LSD para cada nivel de habilidad. En la presente condición se ha evaluado la matriz \mathbf{Q} en conjunto, sin realizar la separación densa-esparcida. Recuerdese también que la matriz \mathbf{Q} no se ha manipulado, los cambios afectan a la matriz de probabilidad $P(A = 1|\theta)$ y consecuentemente a la matriz de probabilidad de acierto de los ítems P_{ij} . Al igual que en las condiciones anteriores se han utilizado los puntos de corte para el estadístico MAD encontrados en el Capítulo 3 para la condición 4A15I y se han señalado con un asterisco los ítems mal recuperados ya que sus valores de MAD exceden los puntos de corte.

Tabla 4.17: Valores MAD para cada ítem y LSD para cada nivel de habilidad

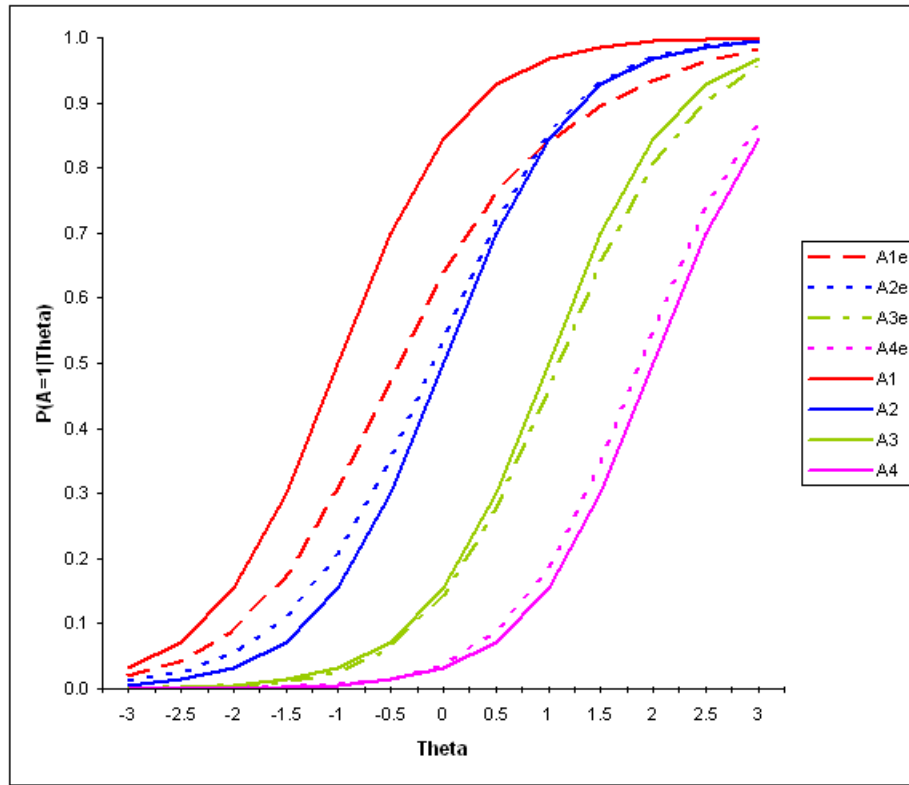
	Q				MAD	θ	LSD
i1	1	0	0	0	0.099*	-3	0.353
i2	0	1	0	0	0.078*	-2.5	0.345
i3	0	0	1	0	0.044	-2	0.331
i4	0	0	0	1	0.011	-1.5	0.313
i5	1	1	0	0	0.038	-1	0.290
i6	1	0	1	0	0.078*	-0.5	0.262
i7	1	0	0	1	0.004	0	0.228
i8	0	1	1	0	0.041	0.5	0.187
i9	0	1	0	1	0.140*	1	0.140
i10	0	0	1	1	0.001	1.5	0.094
i11	0	1	1	1	0.036*	2	0.055
i12	1	0	1	1	0.003	2.5	0.028
i13	1	1	0	1	0.018	3	0.013
i14	1	1	1	0	0.085*		
i15	1	1	1	1	0.027*		

* ítems mal recuperados según punto de corte.

En la Tabla 4.17 se puede observar que la heterogeneidad de los atributos ejerce un efecto importante en la recuperación de las CCIs, pues siete de los quince ítems presentan una recuperación pobre. En la tabla también se puede apreciar que los ítems mal recuperados son los que requieren los dos atributos de mayor heterogeneidad (A_1 y A_2), estos son los dos atributos han sido generados con mayor dispersión en los parámetros de dificultad b . En cuanto a los valores LSD se puede notar que son diferentes de 0 pero no se incrementan tanto como en los escenarios previos.

La Gráfica 4.27 exhibe las curvas de probabilidad estimadas con el LSDM. En esta gráfica se observa que la forma de las curvas es muy similar a la original, excepto para el atributo A_1 . Lo anterior se ve confirmado con los valores RMSE de 0.136 (A_1), 0.026 (A_2), 0.024 (A_3) y 0.024 (A_4). Este resultado era esperable, ya que A_1 es el atributo más heterogéneo.

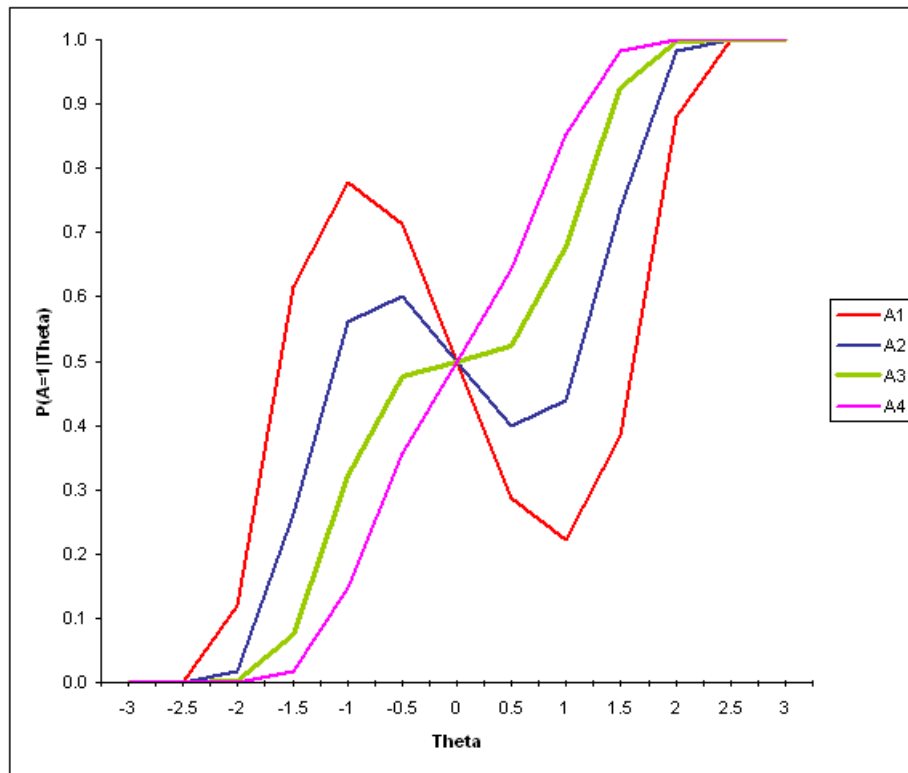
Gráfica 4.27: CPAs para la condición de heterogeneidad



4.3.2.2. Análisis de los efectos de la falta de monotonocidad

La Gráfica 4.28 presenta las CPAs estimadas para esta condición. En la gráfica se puede observar que el modelo reproduce las curvas tal y como fueron generadas; esto se refleja también en los valores MAD que son, en todos los casos, iguales a 0. Este resultado indica que, contrario a lo que se creía, la falta de monotonocidad de los atributos no afecta la precisión de la estimación de las CPAs con el LSDM.

Gráfica 4.28: CPAs para la condición de falta de monotonicidad



4.4. Conclusiones

En el presente trabajo se han introducido cuatro tipos de error de especificación en una matriz \mathbf{Q} ficticia de 4 atributos y 15 ítems con el objeto de examinar sus efectos sobre los criterios de validación del LSDM. En primer lugar, es importante anotar que los errores no han sido introducidos al azar, como en el estudio de Baker (1993), sino que se ha procurado analizar el efecto de errores que pueden ocurrir en un contexto aplicado o en estudios reales. En segundo lugar, se ha intentado tener en cuenta la mayor parte de fuentes de error que se han podido identificar tanto en los estudios sobre los efectos de los errores en \mathbf{Q} (Baker, 1993; Rupp y Templin, 2008; de la Torre, 2008), como en los estudios empíricos revisados. En tercer lugar, hay que tener en cuenta que este estudio es de carácter descriptivo y exploratorio por lo tanto no se pretende realizar inferencias ni generalizaciones.

Los resultados del presente estudio indican que los diversos tipos de errores de especificación producen efectos importantes en los criterios del método LSDM. Los errores en

los vínculos ítem-atributo producen un empeoramiento de la recuperación de los ítems, especialmente cuando se incluyen vínculos irrelevantes en los atributos difíciles. La recuperación de las CCIs también empeora cuando se fusionan atributos difíciles o cuando hay incompatibilidades entre ellos. También se ha detectado una mala recuperación en ítems que requieren atributos heterogéneos cuyas dificultades son diferentes según el ítem en el que se utilicen. Adicionalmente, al igual que en el estudio de de la Torre (2008), algunos ítems correctamente especificados fueron afectados por la manipulación de \mathbf{Q} , en otras palabras, mediante el estadístico MAD, se puede detectar la presencia de errores de especificación pero no se puede conocer el ítem concreto que tiene problemas de especificación.

Los valores residuales (LSD) se han visto incrementados en la mayoría de las condiciones, especialmente en los niveles bajos de habilidad, confirmando lo encontrado en otras aplicaciones del LSDM: el método es más preciso en niveles altos de rasgo. La estimación de las CPAs también es un criterio importante en la detección de errores de especificación. La forma y orden de dificultad de las CPAs suele verse alterada en la mayoría de las condiciones, aunque no ocurre lo mismo con su monotonidad. Los errores de especificación suelen producir sobreestimación de las probabilidades haciendo que las curvas de algunos atributos, especialmente las de los más fáciles, se acerquen a un valor de 1 en todos los niveles de habilidad. La fusión de atributos difíciles, las incompatibilidades y los errores en los vínculos que involucran a los atributos con mayor dificultad relativa producen alteraciones en las CPAs, especialmente en las de los atributos fáciles.

Otro resultado a destacar es que la inclusión de un atributo irrelevante y la falta de monotonidad de los atributos no tienen efectos en la recuperación de los ítems (valores $\text{MAD} = 0$). En el caso de la adición de un atributo irrelevante, se ha encontrado que las probabilidades de dominio del atributo introducido son de uno para todos los niveles de habilidad mientras que los demás valores de probabilidad no se encuentran alterados. Contrario a lo que se esperaba, las relaciones de dependencia secuencial no producen efectos importantes en los criterios de validación del LSDM.

En cuanto a la división entre matrices \mathbf{Q} densas y esparcidas, se observa, de forma similar a lo encontrado por Baker (1993), una mayor influencia de los errores en los vínculos y de fusión entre atributos en las matrices de tipo esparcido. Sin embargo, otros errores no incluidos por dicho autor en su estudio como los errores de incompatibilidad y depen-

dencia tienen un mayor efecto en las matrices densas. Otra conclusión interesante es que ocurre sobrestimación de las probabilidades de dominio de los atributos cuando se incluyen vínculos ítem-atributo no contemplados originalmente en \mathbf{Q} e infraestimación cuando se eliminan vínculos necesarios entre ítems y atributos, esta tendencia es más evidente cuando los atributos afectados son difíciles.

Es notoria la influencia de la heterogeneidad en los criterios del LSDM, el método parte de un supuesto de homogeneidad de la dificultad de los atributos en el conjunto de ítems que los requieren y por lo tanto, cuando el investigador tenga sospechas de que esta condición no se cumple, es mejor tener precaución con los resultados. Dimitrov (2007) propone como criterio de validación de los atributos cognitivos la monotonicidad de las CPAs, sin embargo, en ninguna de las condiciones aquí analizadas se produce una falta de monotonicidad, por lo tanto dicho criterio no parece ser un indicador importante de la presencia de errores de especificación, como sí lo es la infraestimación y, sobre todo, la sobrestimación de las probabilidades de los atributos, de allí que curvas de probabilidad de atributo fijas en uno (a lo largo de los diferentes niveles de habilidad) indican la presencia de errores de especificación de \mathbf{Q} , concretamente, en los atributos difíciles contemplados en la estructura. Con base a estos resultados se puede concluir que, bajo ciertas condiciones, el LSDM es una herramienta útil para acopiar evidencia sobre la validez de la estructura cognitiva en este tipo de contextos. Sin embargo, las condiciones abarcadas son limitadas, no se han probado, por ejemplo, otros tamaños de test ni el número de atributos implicados. Como el objetivo del presente estudio ha sido realizar una descripción y exploración inicial, el siguiente paso lógico es la inferencia mediante la realización de réplicas y el estudio de la sensibilidad del método para la detección de los diversos tipos de error de especificación. Este es por tanto, el objetivo del siguiente estudio (Capítulo 5).

Capítulo 5

Sensibilidad del LSDM para la detección de errores en \mathbf{Q}

5.1. Introducción

En el presente estudio se analiza si los estadísticos MAD y LSD son sensibles a la presencia de errores de especificación de la matriz \mathbf{Q} . En un estudio previo (ver Capítulo 4) se ha evidenciado que los valores de MAD y LSD proporcionan información sobre la calidad de la estructura cognitiva representada en \mathbf{Q} , es por ello que resulta fundamental generalizar sus propiedades respecto a la detección de distintos tipos de errores de especificación y también establecer puntos de corte que permitan al investigador detectar la presencia de error.

El objetivo del presente trabajo es analizar las propiedades estadísticas (sensibilidad y especificidad) de los criterios MAD y LSD para la detección de diferentes tipos de errores de especificación en la matriz \mathbf{Q} en diversas condiciones en las que se manipula el número de ítems y de atributos que componen el test.

5.2. Procedimiento

5.2.1. Diseño del estudio

Las variables independientes contempladas en este trabajo son el número de ítems del test (15 o 32) y el número de atributos (4 u 8); por lo tanto el estudio tienen $2 \times 2 = 4$

condiciones: dos de ellas tienen 4 atributos y 15 ó 32 ítems (4A15I y 4A32I) y las otras dos contienen 8 atributos y 15 ó 32 ítems (8A15I y 8A32I).

5.2.2. Diseño de la simulación

La simulación tiene varios pasos que se describirán a continuación.

5.2.2.1. Generación de las matrices \mathbf{Q}

Se han generado aleatoriamente cuatro matrices \mathbf{Q} (una para cada condición). Dichas matrices se consideran verdaderas y se han usado para introducir los errores de especificación. En el proceso de generación de las matrices \mathbf{Q} se aplicaron ciertas restricciones con el fin de obtener matrices exhaustivas: como las matrices \mathbf{Q} son de carácter binario para todas las condiciones se formaron las posibles combinaciones de 0 y 1 excluyendo la combinación de ceros (pues cada ítem requiere como mínimo un atributo). La matriz \mathbf{Q} correspondiente a la condición 4A15I incluye los patrones de todas las combinaciones binarias posibles ($2^4 - 1 = 15$). Para la condición 4A32I se ha duplicado la matriz \mathbf{Q} de la condición 4A15I y se han incluido dos ítems más seleccionados al azar.

En las condiciones con 8 atributos ha sido necesario emplear un procedimiento de selección de patrones. Este proceso se ha llevado a cabo respetando las proporciones del número de atributos requerido por cada ítem como se expone en la Tabla 5.1, en la que la columna A representa el número de atributos requeridos por el ítem, la columna P el número de patrones posibles que se conforman con la cantidad respectiva de atributos y las dos últimas columnas muestran el número de ítems incluidos en cada condición que se han seleccionado aleatoriamente entre los P patrones posibles. Por ejemplo, en la matriz \mathbf{Q} de la condición 8A32I se han incluido 9 patrones seleccionados al azar de los 70 patrones posibles que se conforman cuando se combinan 4 atributos. El Anexo 2 contiene las cuatro matrices \mathbf{Q} *verdaderas* que se han usado como punto de partida para la introducción de errores de especificación.

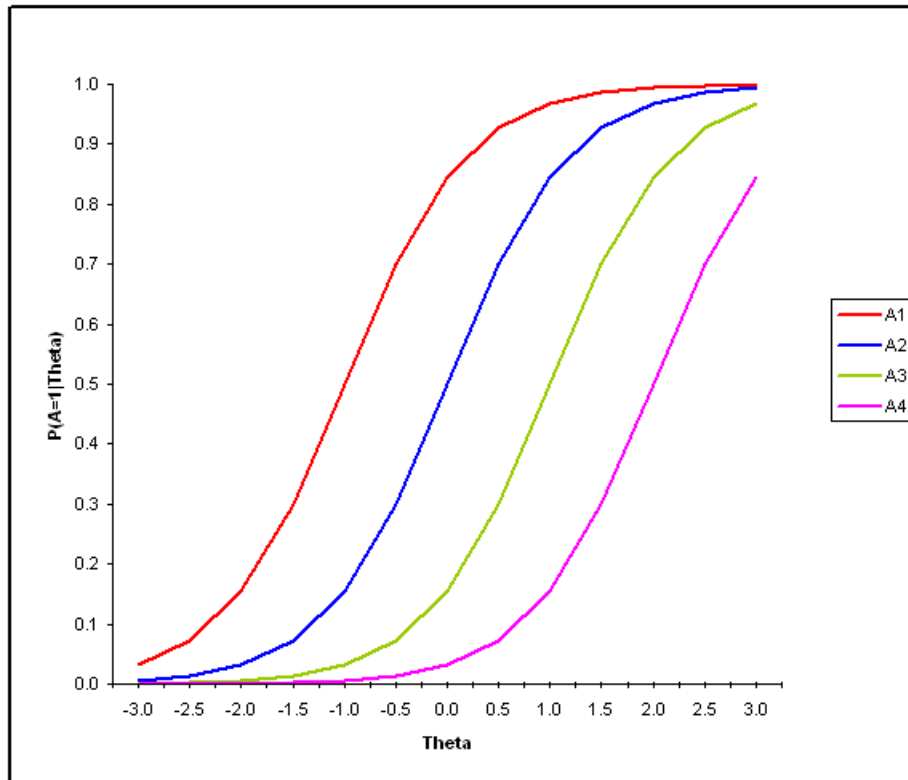
Tabla 5.1: Distribución de ítems en las matrices **Q** según el número posible de patrones

A	P	8A15I	8A32I
1	8	1	2
2	28	1	3
3	56	3	6
4	70	4	9
5	56	3	6
6	28	1	3
7	8	1	2
8	1	1	1
Total	255	15	32

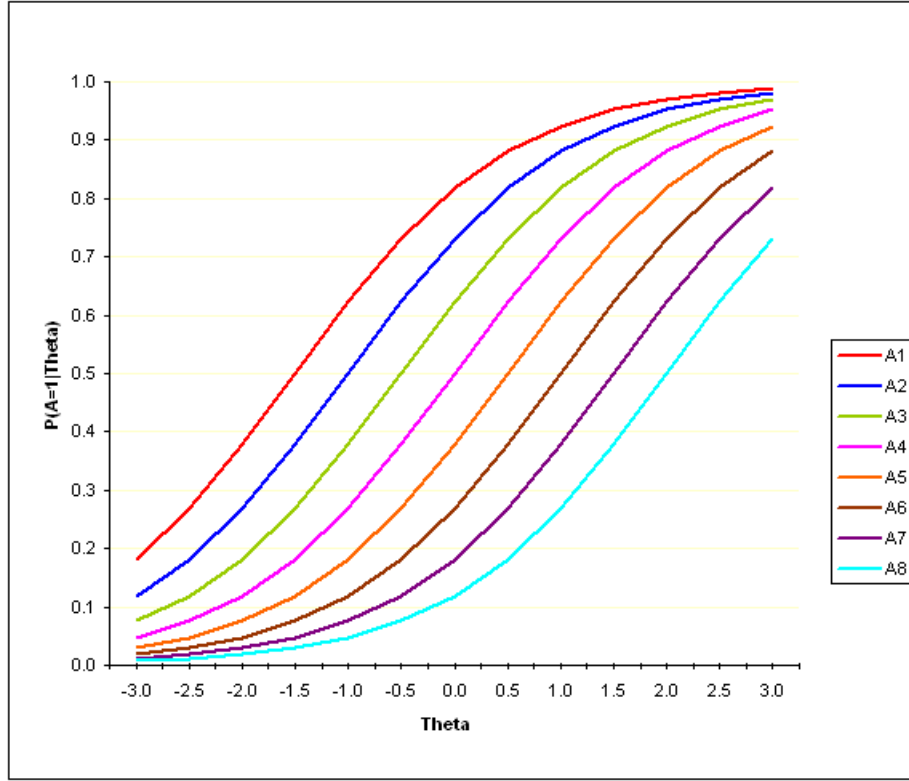
5.2.2.2. Generación de las matrices: $P(A = 1|\theta)$

En este paso se han simulado las probabilidades de dominio de los atributos utilizando el modelo de Rasch con parámetros de dificultad (b) de: -1 (A_1), 0 (A_2), 1 (A_3) y 2 (A_4) para las condiciones con cuatro atributos y de -1.5 (A_1), -1 (A_2), -0.5 (A_3), 0 (A_4), 0.5 (A_5), 1 (A_6) 1.5 (A_7) y 2 (A_8) para las condiciones con ocho atributos. De esta forma se obtienen atributos con una dificultad creciente como se puede apreciar en las Gráficas 5.1 y 5.2.

Gráfica 5.1: CPAs verdaderas para las condiciones con 4 atributos



Gráfica 5.2: CPAs verdaderas para las condiciones con 8 atributos



5.2.2.3. Generación de matrices P_{ij}

La probabilidad de resolver correctamente el ítem j en cada nivel fijo de habilidad θ_i , (P_{ij}) , ha sido calculada en cada condición utilizando la ecuación básica del LSDM, tomando las probabilidades $P(A = 1|\theta)$ halladas en el paso anterior y los valores q_{ik} de las matrices \mathbf{Q} verdaderas:

$$P_{ij} = \prod_{k=1}^K [P(A = 1|\theta)]^{q_{ik}} \quad (5.1)$$

5.2.2.4. Verificación de la calidad técnica del LSDM

Antes de ejecutar el presente estudio se ha verificado la ejecución técnica del LSDM usando como entrada al procedimiento las matrices P_{ij} verdaderas, resultantes en el paso anterior, y se ha encontrado, como se esperaba, una perfecta recuperación de las CPAs (RMSE = 0 en todas las condiciones), de las CCIIs (MAD = 0 en todos los ítems en todas las condiciones) y LSD = 0 en todas las condiciones, por lo cual se garantiza un adecuado punto de partida para el estudio de las consecuencias de los errores de especificación.

5.2.2.5. Obtención de réplicas de las matrices P_{ij}

Se han realizado 200 réplicas de cada una de las cuatro matrices P_{ij} , correspondientes a cada condición. Para ello se ha simulado una variable aleatoria de distribución binomial con parámetros p iguales a cada uno de los elementos de la matriz de probabilidad $p = P_{ij}$ y $n = 1000000$. De acuerdo a dicho procedimiento, un elemento particular de una réplica de P_{ij} corresponde al valor obtenido de la distribución binomial dividido en un millón (n). Dicho procedimiento se ha repetido 200 veces para obtener el mismo número de réplicas aleatorias de P_{ij} para cada condición. Cada réplica de P_{ij} resulta ligeramente diferente a la original, de forma análoga a la técnica de bootstrapping (Mooney y Duval, 1993), con la diferencia de que la matriz sobre la que se toman las réplicas es la matriz P_{ij} , en lugar de realizar las re-muestras sobre una matriz de datos, que es como se aplica tradicionalmente dicha técnica.

En el presente estudio se han realizado réplicas de la matriz P_{ij} , en lugar de replicar matrices de respuestas a los ítems, porque el objetivo es evaluar la sensibilidad y especificidad de los criterios del LSDM cuando se cumple la condición de que las probabilidades de responder correctamente los ítems son el producto de las probabilidades de dominio en los atributos requeridos.

5.2.2.6. Introducción de errores de especificación en las matrices \mathbf{Q}

Se han introducido 14 tipos de error en las matrices \mathbf{Q} generadas en el primer paso, procurando no producir un vector fila de ceros en ningún caso. En los seis tipos de error que se expondrán a continuación se manipulan únicamente los atributos de mayor dificultad, A_3 y A_4 en el caso de 4 atributos y A_5 , A_6 , A_7 y A_8 en el de 8 atributos:

1. Dificiles-Incluir (DI): este tipo de error consiste en cambiar ceros por unos (incluir vínculos) en los atributos de mayor dificultad relativa.
2. Dificiles-Excluir (DE): este tipo de error consiste en cambiar unos por ceros (excluir vínculos) en los atributos de mayor dificultad relativa.
3. Dificiles-AND-Fusionar (DAF): los dos atributos con mayor dificultad relativa se fusionan siguiendo la regla *AND*, es decir, un elemento particular fusionado sólo asume

un valor de cero cuando los dos elementos originales son cero, en caso contrario, asume un valor de uno.

4. Dificiles-OR-Fusionar (DOF): los dos atributos con mayor dificultad se fusionan siguiendo la regla *OR*, es decir, un elemento particular fusionado sólo asume el valor uno cuando los dos elementos originales son uno, en caso contrario, asume el valor de cero.
5. Dificiles-Dependencia Secuencial (DDS): la dependencia secuencial representa la relación entre un par de atributos según la cual, si uno de los atributos es requerido para la respuesta correcta del ítem el otro también debe estar presente, en este caso la dependencia ocurre entre los atributos más difíciles. En las condiciones de 4 atributos, A_3 depende de A_4 ; en otras palabras, si $q_{i4} = 1$, entonces $q_{i3} = 1$. En las condiciones con 8 atributos A_5 depende de A_6 , y A_7 depende de A_8 .
6. Dificiles-Incompatibles (DIN): La incompatibilidad entre un par de atributos consiste en que si un atributo es requerido para la correcta solución del ítem el otro no puede estar presente. De nuevo, los errores han sido introducidos en parejas que involucran los atributos más difíciles.

En la Tabla 5.2 se presenta, a modo de ilustración, la matriz **Q** verdadera y las 6 matrices **Q** con errores de especificación introducidos en los atributos difíciles, para la condición 4A15I. En dicha tabla sólo se han incluido las columnas correspondientes a los atributos difíciles (que han sido objeto de algún tipo de modificación) y los elementos cambiados se resaltan en negrilla.

Tabla 5.2: Algunas matrices \mathbf{Q} con errores de especificación de la condición 4A15I

Q verdadera				DI		DE		DAM	DOM	DDS		DIN	
A_1	A_2	A_3	A_4	A_3	A_4	A_3	A_4	A_{3-4}	A_{3-4}	A_3	A_4	A_3	A_4
1	0	0	0	0	1	0	0	0	0	0	0	0	0
0	1	0	0	0	1	0	0	0	0	0	0	0	0
0	0	1	0	1	1	1	0	1	0	1	0	1	0
0	0	0	1	1	1	0	1	1	0	1	1	0	1
1	1	0	0	1	0	0	0	0	0	0	0	0	0
1	0	1	0	1	1	0	0	1	0	1	0	1	0
1	0	0	1	1	1	0	0	1	0	1	1	0	1
0	1	1	0	1	1	0	0	1	0	1	0	1	0
0	1	0	1	1	1	0	0	1	0	1	1	0	1
0	0	1	1	1	1	0	1	1	1	1	1	0	1
0	1	1	1	1	1	0	1	1	1	1	1	0	1
1	0	1	1	1	1	0	1	1	1	1	1	0	1
1	1	0	1	1	1	0	0	1	0	1	1	0	1
1	1	1	0	1	1	0	0	1	0	1	0	1	0
1	1	1	1	1	1	1	0	1	1	1	1	0	1

Los siguientes seis tipos de error de especificación son iguales a los ya explicados, pero se aplican únicamente en atributos con baja dificultad, es decir, A_1 y A_2 en el escenario de 4 atributos y A_1 , A_2 , A_3 y A_4 en las condiciones con 8 atributos.

1. Fáciles-Incluir (FI): este tipo de error consiste en cambiar ceros por unos (incluir vínculos) en los atributos de menor dificultad relativa.
2. Fáciles-Excluir (FE): este tipo de error consiste en cambiar unos por ceros (excluir vínculos) en los atributos de menor dificultad relativa.
3. Fáciles-AND-Fusionar (FAF): los dos atributos con menor dificultad relativa se fusionan siguiendo la regla *AND*, es decir, un elemento particular fusionado sólo asume un valor de cero cuando los dos elementos originales son cero, en caso contrario, asume un valor de uno.
4. Fáciles-OR-Fusionar (DOF): los dos atributos con menor dificultad relativa se fusionan siguiendo la regla *OR*, es decir, un elemento particular fusionado sólo asume un valor de uno cuando los dos elementos originales son uno, en caso contrario, asume un valor de cero.
5. Fáciles-Dependencia Secuencial (FDS): la dependencia secuencial representa la relación entre un par de atributos según la cual, si uno de los atributos es requerido para

la respuesta correcta del ítem, el otro también debe estar presente, en este caso la dependencia ocurre entre los atributos más fáciles. En las condiciones de 4 atributos, A_1 depende de A_2 ; en otras palabras, si $q_{i2} = 1$, entonces $q_{i1} = 1$. En las condiciones con 8 atributos A_1 depende de A_2 , y A_3 depende de A_4 .

6. Fáciles-Incompatibles (FIN): La incompatibilidad entre un par de atributos consiste en que, si un atributo es requerido para la correcta solución del ítem el otro no puede estar presente. De nuevo, los errores han sido introducidos en parejas que involucran los atributos más fáciles, así, A_1 es incompatible con A_2 y A_3 con A_4 .

Los dos últimos tipos de error de especificación afectan a toda la matriz \mathbf{Q} .

1. Adicionar (AD): consiste en adicionar un atributo irrelevante a la matriz \mathbf{Q} , las entradas de dicho atributo han sido seleccionadas al azar pero se ha buscado mantener la misma proporción de unos y ceros que en los otros atributos.
2. Dependencia Secuencial Compleja (DSC): se trata de generar dependencias secuenciales entre todos los atributos. Por ejemplo, en las condiciones con 4 atributos A_1 depende de A_2 y este atributo a su vez depende de A_3 y A_4 ; por lo tanto, si $q_{i3} = 1$ o si $q_{i4} = 1$ entonces $q_{i2} = 1$ y, si $q_{i2} = 1$ entonces $q_{i1} = 1$. En el caso de las condiciones con 8 atributos, A_3 depende de A_5 y A_6 , mientras que A_4 depende de A_7 y A_8 . A su vez, A_2 depende de A_3 y A_4 y finalmente A_1 depende de A_2 , en otras palabras, A_1 se requerirá siempre que cualquier otro atributo sea necesario para resolver el ítem, pues la adjudicación de valores es la misma que ya se explicó para el caso de cuatro atributos.

5.2.2.7. Aplicación del método LSDM

Finalmente, se ha ejecutado el LSDM para obtener los valores MAD y LSD, cuyas propiedades se analizan posteriormente mediante la metodología ROC. En cada condición se realizan 200 análisis con el LSDM: uno para cada remuestra de P_{ij} con cada una de las matrices \mathbf{Q} erróneas. Así, en las condiciones 4A15I y 8A15I se han obtenido 200 distribuciones de 15 valores MAD por cada una de las 14 matrices \mathbf{Q} , mientras que en las condiciones 4A32I y 8A32I se han obtenido 200 distribuciones de 32 valores MAD en las 14

condiciones de error. En el caso del criterio LSD se han obtenido 200 distribuciones de 13 valores LSD (recuérdese que hay un valor LSD por cada nivel fijo de habilidad) para cada una de las cuatro condiciones de diseño y para cada una de las 14 matrices \mathbf{Q} . También se ha aplicado el LSDM (usando como entrada cada una de las 200 P_{ij}) a la matriz \mathbf{Q} verdadera y sus resultados se han usado, en conjunto con cada uno de los resultados de las \mathbf{Q} erróneas para ejecutar los análisis ROC y analizar de esta forma la sensibilidad y especificidad de los estadísticos.

5.2.3. Métodos de análisis estadístico

5.2.3.1. Curvas ROC

Para evaluar la especificidad y sensibilidad de los estadísticos MAD y LSD en cada una de las condiciones de error de \mathbf{Q} se ha utilizado la metodología de curvas ROC. Dicha técnica estadística permite organizar, visualizar y analizar la ejecución de clasificadores por medio de un análisis gráfico (Fawcett, 2006). También permite establecer puntos de corte óptimos para la clasificación dicotómica del estado de un fenómeno, en este caso, la correcta o incorrecta recuperación de las CCIs para el caso de los valores MAD y la precisión o imprecisión de la estimación de las CPAs en el caso de los valores LSD. La metodología de curvas ROC también proporciona información sobre la precisión de la clasificación (Franco y Vivo, 2007). Una de las ventajas de esta aproximación es que es útil cuando la distribución del clasificador es asimétrica, como es el caso de los estadísticos MAD y LSD (véase Capítulo 3).

En el proceso de clasificación de ítems de acuerdo a los valores MAD pueden ocurrir cuatro consecuencias: a) si el ítem ha sido correctamente especificado y ha sido clasificado como bien recuperado se denomina verdadero positivo, b) si el mismo ítem es clasificado como mal recuperado se denomina falso negativo, c) si el ítem ha sido incorrectamente especificado y ha sido clasificado en el grupo de ítems bien recuperados se denomina falso positivo y d) si ese mismo ítem es clasificado en el grupo de ítems mal recuperados se denomina verdadero negativo. Basados en estas consecuencias se puede construir una tabla de contingencia, también llamada tabla de confusión, representada en la Tabla 5.3. Dicha tabla contiene la distribución de los ítems en las cuatro posibles consecuencias, los ítems

Tabla 5.3: Tabla de confusión para el estadístico MAD

	Especificación correcta	Especificación incorrecta
Bien recuperado	Verdaderos Positivos (VP)	Falsos Positivos (FP)
Mal recuperado	Falsos Negativos (FN)	Verdaderos Negativos (VN)

Tabla 5.4: Matriz de fracciones

	Especificación correcta	Especificación incorrecta
Bien recuperado	FVP = VP/TQV	FFP = FP/TQF
Mal recuperado	FFN = FN/TQV	FVN = VN/TQF

correctamente clasificados se ubicarán en la diagonal de la tabla, por lo tanto, si la frecuencia de casos en la diagonal es alta, comparada con el total de casos, es posible afirmar que el estadístico MAD es preciso, sin embargo, para evitar juicios subjetivos, es posible estimar dos probabilidades para determinar la precisión del clasificador: la sensibilidad y especificidad.

La sensibilidad (S) es la probabilidad de recuperar adecuadamente la CCI de un ítem que proviene de una matriz \mathbf{Q} correctamente especificada. Su estimación es la proporción de respuestas positivas que son correctamente clasificadas:

$$S = \frac{VP}{VP + FN} = FVP \quad (5.2)$$

También se conoce como Fracción de Verdaderos Positivos (FVP) o exactitud positiva (ver Tabla 5.4).

La especificidad (E) es la probabilidad de recuperar mal la CCI de un ítem que proviene de una \mathbf{Q} mal especificada, el estimador de la especificidad es la proporción de respuestas negativas que son correctamente clasificadas.

$$E = \frac{VN}{FP + VN} = FVN \quad (5.3)$$

También se conoce como Fracción de Verdaderos Negativos (FVN) o exactitud negativa (Tabla 5.3).

En la Tabla 5.4 se presenta la matriz de fracciones, la Fracción de Verdaderos Positivos (FVP) es el número de Verdaderos Positivos (VP) sobre el total de valores provenientes de la \mathbf{Q} verdadera (TQV), la Fracción de Falsos Negativos (FFN) corresponde al número

de Falsos Negativos (FN) sobre el total de valores provenientes de la \mathbf{Q} verdadera (TQV), la Fracción de Falsos Positivos (FFP) es el número de Falsos Positivos (FP) sobre el total de valores provenientes de la \mathbf{Q} falsa y la Fracción de Verdaderos Negativos (FVN) es el número de Verdaderos Negativos sobre el total de valores provenientes de la \mathbf{Q} falsa.

Un clasificador ideal no tendría errores de clasificación ($FN = FP = 0$) y sus valores E y S serían 1. La sensibilidad y especificidad son estimaciones muestrales de las probabilidades poblacionales, por dicho motivo es posible construir intervalos confidenciales basados en su desviación estándar como se puede ver en las siguientes expresiones (Franco y Vivo, 2007):

$$SD(S) = \sqrt{\frac{VP \cdot FN}{(VP + FN)^3}} \quad (5.4)$$

$$SD(E) = \sqrt{\frac{FP \cdot VN}{(FP + VN)^3}} \quad (5.5)$$

Los intervalos de confianza, de S y E respectivamente, se definen como: $S \pm Z_{1-\alpha/2}SD(S)$ y $E \pm Z_{1-\alpha/2}SD(E)$.

Las medidas de precisión expuestas hasta el momento corresponden a un clasificador binario con un solo punto de corte para la división de los grupos, por lo tanto existe una sola pareja de valores S y E para reflejar la precisión del clasificador. Sin embargo, si el estadístico es ordinal o continuo (como el caso que nos ocupa) es posible obtener diversos valores de especificidad y sensibilidad porque en cada punto de corte hay una pareja (E, S) asociada para reflejar la adecuación del clasificador. La idea del procedimiento aquí empleado es buscar el punto de corte óptimo para obtener la clasificación más precisa. En el caso de un clasificador continuo como el MAD o el LSD, la selección del punto de corte afecta la sensibilidad y especificidad. Por ejemplo, si el punto de corte se mueve a la derecha incrementará la sensibilidad pero si se mueve a la izquierda incrementará la especificidad (Franco y Vivo, 2007). En este caso, el punto de corte es definido como el valor a partir del cual el ítem es clasificado en alguno de los dos grupos (correcta o incorrectamente recuperado). Esta situación es análoga a las probabilidades de los errores tipo I y II (α y β respectivamente) en el contexto de las pruebas de hipótesis (Franco y Vivo, 2007).

La curva ROC constituye una representación gráfica del error de tipo I (α) y la potencia estadística ($1 - \beta$) del clasificador. En la curva ROC se representa: a) cada uno de los posibles puntos de corte del clasificador y b) las medidas de precisión de la respectiva clasificación: el error de tipo I ($\alpha = 1 - E$) en el eje X y la sensibilidad ($S = 1 - \beta$) en el eje Y; en otras palabras, la curva ROC representa gráficamente la precisión global del clasificador. En el punto de corte óptimo las dos probabilidades de error son mínimas, sin embargo, si α disminuye β se incrementa y viceversa, por dicho motivo, encontrar puntos de corte adecuados requiere el análisis de la compensación entre el error de tipo I (α) y la potencia ($1 - \beta$). Si un estadístico es deficiente para los propósitos de clasificación no tiene sentido buscar los puntos de corte óptimo.

Existen métodos tanto paramétricos como no paramétricos para calcular las curvas ROC. Los métodos no paramétricos, que han sido los utilizados en el presente estudio, son apropiados cuando no se asume distribución para el clasificador en ninguno de los dos grupos (presencia o ausencia de la condición de interés). El método no paramétrico más común es el empírico (utilizado en este trabajo). Por medio de dicho método se estiman las funciones de distribución empíricas para los grupos de presencia (F_p) y ausencia (F_a). La función de distribución empírica se define como el porcentaje observado de valores muestrales iguales o menores a x para todo punto de corte x . Dicha función es determinada por los pares $[1 - \hat{F}_a(x), 1 - \hat{F}_p(x)]$.

Para cada punto de corte, x ,

$$F_p(x) = \frac{1}{n_p} \sum_{i=1}^{n_p} I(x_i \leq x) \quad (5.6)$$

y

$$F_a(x) = \frac{1}{n_a} \sum_{j=1}^{n_a} I(x_j \leq x) \quad (5.7)$$

Donde I es una función indicadora que vale 1 cuando x_i o x_j son inferiores al punto de corte. En otro caso, vale 0.

Conectando estos pares de puntos por líneas se obtiene la curva ROC empírica, que se suaviza cuando el clasificador es continuo, tendiendo a converger con la curva ROC teórica. Según Franco y Vivo (2007), la estimación con métodos no paramétricos tiene

algunas ventajas: a) utiliza toda la información disponible pues la curva pasa por todos los puntos de corte, b) proporciona estimaciones insesgadas de sensibilidad, especificidad y área bajo la curva porque en cada punto de corte el mejor estimador es el no-paramétrico, c) no requiere el cumplimiento de supuestos y d) sus cálculos son simples. Este método también presenta algunas desventajas como: a) la forma escalonada de la curva, b) los empates pueden causar infraestimación de las curvas y c) no permite comparar los puntos de dos curvas.

Para obtener una visión general de la precisión del clasificador y para realizar comparaciones entre clasificadores es necesario evaluar el área bajo la curva ROC (en inglés, AUC). El AUC asume valores entre 0 y 1, donde 1 corresponde a la mayor precisión del clasificador (discriminación perfecta entre los dos estados) y 0 representa la situación opuesta, es decir, una discriminación perfecta pero en el sentido equivocado. Un AUC de 0.5 representa una clasificación aleatoria y la tasa VP será igual a la de FP. Los valores AUC deben estar entre 0.5 y 1 y cuanto más cercanos a 1 mucho mejor (Franco y Vivo, 2007). La comparación entre dos clasificadores no debe basarse exclusivamente en el AUC porque pueden existir dos curvas ROC muy diferentes pero con AUC bastante similar. El uso del área parcial bajo la curva (PAUC) permite manejar correctamente este tipo de situación.

Existen situaciones en las que una medida general es inadecuada por las características del clasificador, en dichos casos es necesario obtener una medida de la calidad del mismo entre valores parciales de falsos positivos. El PAUC es una porción de AUC restringida a una parte del espacio ROC, dicha área se suele estandarizar para poderla interpretar de la misma forma que se hace con el área total, de nuevo, valores entre 0.5 y 1 son aceptables, especialmente cuando se acercan a 1 (Franco y Vivo, 2007). El PAUC también se puede interpretar como la sensibilidad promedio en el intervalo seleccionado de valores de especificidad.

5.2.3.2. Análisis ROC de los valores LSDM y MAD

Con base a los resultados de la aplicación del LSDM (paso previo) se generaron tablas 2x2 usando los valores MAD (y LSD) de la condición verdadera y de cada condición de error particular. Los estadísticos ROC: punto de corte para el estadístico a una FFP =

0.05, FVP o sensibilidad, y área parcial (PAUC) fueron obtenidos para cada tabla. Dicho procedimiento se repitió para cada una de las 200 réplicas en cada condición (4A15I, 4A32I, 8A15I y 8A32I).

Los estadísticos ROC que se presentan en los resultados del presente estudio son **medias** de los estadísticos ROC de cada una de las 200 réplicas analizadas, dichos estadísticos son:

- a) Valor promedio de los puntos de corte para el MAD y LSD asociados a una FFP = 0.05.
- b) Valor promedio de la FVP o sensibilidad (S) de los criterios MAD y LSD.
- c) Área parcial bajo la curva ROC (PAUC) **estandarizada** para valores de FFP entre 0 y 0.05.
- d) Intervalos de confianza al 95 % para la **media** de los tres estadísticos anteriormente mencionados.
- e) Desviación estándar de la distribución de los estadísticos ROC.

Todos los procedimientos utilizados tanto en la simulación como en el análisis estadístico han sido programados en rutinas en el entorno R.

5.3. Resultados

En las Tablas 5.5 y 5.6 se presentan los estadísticos ROC de los criterios MAD y LSD para la condición 4A15I. En dichas tablas se presenta el promedio de los valores MAD asociados a una FFP de 0.05 con sus respectivos intervalos de confianza al 95 % y su desviación estándar (sd), dicho promedio es el punto de corte que se ha considerado para la estimación del área bajo la curva parcial estandarizada, existiendo un punto de corte diferente para cada análisis ROC según la condición de error de **Q** (recuérdese que los análisis son independientes). También se reportan la Fracción de Verdaderos Positivos (FVP) y el área parcial estandarizada PAUC media con su respectivo intervalo de confianza al 95 % y desviación estándar (sd).

Tabla 5.5: Estadísticos ROC del criterio MAD en la condición 4A15I

Q	MAD* (CI 95 %)	<i>sd</i>	FVP (CI 95 %)	<i>sd</i>	PAUC (CI 95 %)
AD	0.064 (0.061-0.066)	0.044	0.108 (0.105-0.110)	0.047	0.068 (0.064-0.071)
DE	0.108 (0.105-0.110)	0.047	0.875 (0.872-0.877)	0.040	0.712 (0.706-0.727)
DOF	0.041 (0.039-0.042)	0.024	0.865 (0.864-0.865)	0.008	0.763 (0.754-0.772)
DI	0.040 (0.037-0.040)	0.018	0.982 (0.980-0.983)	0.029	0.837(0.824-0.850)
DAF	0.043 (0.041-0.044)	0.022	0.900 (0.896-0.903)	0.060	0.671 (0.656-0.686)
DIN	0.047 (0.046-0.048)	0.019	0.982 (0.980-0.983)	0.029	0.576 (0.566-0.586)
DDS	0.061 (0.060-0.062)	0.021	0.308 (0.304-0.311)	0.061	0.198 (0.192-0.204)
FE	0.057 (0.055-0.059)	0.028	0.374 (0.368-0.379)	0.093	0.199 (0.190-0.208)
FOF	0.208 (0.197-0.218)	0.159	0.344 (0.340-0.347)	0.049	0.233 (0.227-0.240)
FI	0.005 (0.005-0.005)	0.003	0.410 (0.405-0.414)	0.069	0.239 (0.230-0.247)
FAF	0.051 (0.049-0.053)	0.026	0.330 (0.326-0.333)	0.054	0.213 (0.210-0.220)
FIN	0.062 (0.061-0.063)	0.021	0.310 (0.306-0.313)	0.062	0.198 (0.192-0.203)
FDS	0.061 (0.059-0.062)	0.021	0.312 (0.308-0.315)	0.061	0.199 (0.194-0.205)
DSC	0.042 (0.041-0.043)	0.016	0.390 (0.384-0.395)	0.092	0.208 (0.199-0.208)

* Valor promedio del estadístico MAD asociado a una FFP de 0.05

En la Tabla 5.5 se puede apreciar que el criterio MAD es sensible a errores de especificación introducidos en los atributos de mayor dificultad (excepto para la condición de dependencia secuencial) con FVPs que oscilan entre 0.87 y 0.98. La FVP de los otros tipos de error introducidos en \mathbf{Q} revelan que la clasificación es muy imprecisa ya que el porcentaje de ítems que son clasificados como bien recuperados proviniendo de la \mathbf{Q} verdadera es muy bajo. Los valores del área parcial bajo la curva mayores a 0.5 confirman esta tendencia indicando una buena precisión del criterio MAD, pero únicamente en las condiciones DE, DOF, DI, DAF y DIN. Los puntos de corte (valores MAD a una $FFP = 0.05$) indican que, por encima de 0.037, se detecta con precisión la presencia de errores de especificación en los atributos difíciles.

Tabla 5.6: Estadísticos ROC del criterio LSD en la condición 4A15I

Q	LSD (CI 95 %)	<i>sd</i>	FVP (CI 95 %)	<i>sd</i>	PAUC (CI 95 %)
AD	0.672 (0.666-0.677)	0.090	0.090 (0.086-0.093)	0.057	0.046 (0.040-0.052)
DE	0.809 (0.801-0.817)	0.123	0.346 (0.343-0.348)	0.044	0.304 (0.297-0.310)
DOF	0.801 (0.790-0.811)	0.162	0.416 (0.413-0.418)	0.043	0.343 (0.336-0.350)
DI	0.880 (0.875-0.884)	0.064	0.361 (0.358-0.363)	0.041	0.332 (0.326-0.337)
DAF	0.812 (0.807-0.816)	0.073	0.340 (0.337-0.342)	0.040	0.308 (0.301-0.316)
DIN	0.798 (0.793-0.802)	0.077	0.328 (0.325-0.330)	0.041	0.302 (0.294-0.309)
DDS	0.745 (0.741-0.748)	0.061	0.211 (0.206-0.215)	0.075	0.136 (0.127-0.145)
FE	0.865 (0.855-0.874)	0.152	0.190 (0.185-0.194)	0.065	0.136 (0.130-0.142)
FOF	0.878 (0.866-0.889)	0.183	0.228 (0.223-0.232)	0.067	0.159 (0.122-0.165)
FI	0.741 (0.736-0.745)	0.067	0.238 (0.232-0.243)	0.082	0.158 (0.148-0.168)
FAF	0.749 (0.745-0.752)	0.063	0.221 (0.216-0.225)	0.076	0.146 (0.137-0.155)
FIN	0.739 (0.734-0.743)	0.064	0.214 (0.209-0.218)	0.072	0.139 (0.129-0.148)
FDS	0.753 (0.749-0.756)	0.054	0.207 (0.202-0.211)	0.071	0.133 (0.125-0.141)
DSC	0.773 (0.766-0.779)	0.104	0.291 (0.286-0.295)	0.076	0.220 (0.210-0.229)

* Valor promedio del estadístico LSD asociado a una FFP de 0.05

En la Tabla 5.6 se puede ver que los valores del estadístico residual LSD no son sensibles para la detección de ninguno de los tipos de error introducidos en la estructura cognitiva, pues la FVP está por debajo de 0.41 indicando una clasificación aleatoria entre el modelo mal especificado y el correcto; además, en todos los casos los valores PAUC son inferiores a 0.5. Una explicación a dicho resultado es que los valores LSD tienen un tamaño decreciente, siendo altos en los niveles bajos de habilidad y bajos en los niveles altos de habilidad. Por lo cual, para encontrar un punto de corte apropiado, parece más indicado utilizar únicamente el valor LSD máximo (correspondiente al nivel $\theta = -3$) en lugar del valor asociado a la FFP de 0.05 (como se ha realizado en el presente estudio).

En las Tablas 5.7 y 5.8 se presentan los estadísticos ROC de los criterios MAD y LSD para la condición 4A32I en cada una de las condiciones de error. Los resultados son muy similares a los encontrados en la condición previa, con FVPs entre 0.72 y 0.99 que indican que entre un 72 % y 99 % de los ítems provenientes de la **Q** verdadera se clasifican correctamente en el grupo de ítems bien recuperados. También se han encontrado PAUCs mayores a 0.61 en las citadas condiciones (DE, DOF, DI, DAF y DIN). Los puntos de corte para el estadístico MAD son estables e indican que valores por encima de 0.041 revelan la presencia de este tipo de errores de especificación. Al igual que en la condición anterior, los valores LSD no son sensibles a los errores de especificación de **Q**.

Tabla 5.7: Estadísticos ROC del criterio MAD en la condición 4A32I

Q	MAD* (CI 95 %)	<i>sd</i>	FVP (CI 95 %)	<i>sd</i>	PAUC (CI 95 %)
AD	0.043 (0.042-0.044)	0.016	0.082 (0.079-0.084)	0.034	0.053 (0.050-0.056)
DE	0.044 (0.043-0.045)	0.016	0.920 (0.918-0.921)	0.028	0.709 (0.699-0.718)
DOF	0.043 (0.041-0.046)	0.017	0.876 (0.875-0.876)	0.011	0.746 (0.736-0.756)
DI	0.043 (0.042-0.044)	0.017	0.988 (0.986-0.989)	0.023	0.795 (0.783-0.807)
DAF	0.044 (0.042-0.044)	0.016	0.999 (0.998-0.999)	0.001	0.714 (0.700-0.728)
DIN	0.044 (0.043-0.045)	0.016	0.719 (0.718-0.719)	0.005	0.611 (0.600-0.623)
DDS	0.043 (0.043-0.045)	0.016	0.421 (0.414-0.427)	0.105	0.235 (0.229-0.242)
FE	0.043 (0.042-0.044)	0.016	0.401 (0.396-0.405)	0.076	0.201 (0.194-0.208)
FOF	0.043 (0.042-0.044)	0.016	0.353 (0.349-0.356)	0.048	0.217 (0.213-0.222)
FI	0.044 (0.043-0.045)	0.016	0.417 (0.411-0.422)	0.085	0.261 (0.254-0.267)
FAF	0.043 (0.042-0.044)	0.016	0.386 (0.379-0.392)	0.106	0.219 (0.211-0.226)
FIN	0.043 (0.042-0.044)	0.016	0.394 (0.389-0.398)	0.076	0.229 (0.222-0.235)
FDS	0.044 (0.043-0.045)	0.016	0.186 (0.185-0.187)	0.010	0.088 (0.085-0.092)
DSC	0.043 (0.042-0.044)	0.016	0.393 (0.388-0.397)	0.077	0.182 (0.175-0.189)

* Valor promedio del estadístico MAD asociado a una FFP de 0.05

Tabla 5.8: Estadísticos ROC del criterio LSD en la condición 4A32I

Q	LSD (CI 95 %)	<i>sd</i>	FVP (CI 95 %)	<i>sd</i>	PAUC (CI 95 %)
AD	0.569 (0.567-0.570)	0.029	0.155 (0.152-0.157)	0.041	0.074 (0.069-0.079)
DE	0.669 (0.667-0.670)	0.031	0.368 (0.365-0.370)	0.032	0.321 (0.320-0.322)
DOF	0.647 (0.644-0.649)	0.045	0.384 (0.381-0.386)	0.039	0.324 (0.323-0.326)
DI	0.626 (0.621-0.630)	0.066	0.399 (0.396-0.401)	0.041	0.354 (0.350-0.358)
DAF	0.597 (0.594-0.599)	0.036	0.394 (0.391-0.396)	0.040	0.372 (0.368-0.376)
DIN	0.608 (0.604-0.611)	0.055	0.399 (0.397-0.400)	0.030	0.302 (0.294-0.309)
DDS	0.625 (0.624-0.625)	0.006	0.379 (0.374-0.383)	0.075	0.323 (0.322-0.325)
FE	0.622 (0.617-0.626)	0.076	0.336 (0.331-0.340)	0.075	0.235 (0.230-0.239)
FOF	0.620 (0.615-0.624)	0.067	0.355 (0.350-0.359)	0.067	0.159 (0.122-0.165)
FI	0.583 (0.581-0.584)	0.027	0.238 (0.232-0.243)	0.082	0.290 (0.284-0.296)
FAF	0.588 (0.586-0.589)	0.030	0.377 (0.374-0.379)	0.043	0.284 (0.277-0.290)
FIN	0.588 (0.585-0.590)	0.037	0.366 (0.363-0.368)	0.035	0.288 (0.282-0.295)
FDS	0.627 (0.626-0.627)	0.006	0.386 (0.383-0.388)	0.040	0.325 (0.323-0.326)
DSC	0.610 (0.608-0.611)	0.026	0.379 (0.377-0.380)	0.031	0.316 (0.313-0.319)

* Valor promedio del estadístico LSD asociado a una FFP de 0.05

En las Tablas 5.9 y 5.10 se exhiben los resultados para la condición 8A15I. En la Tabla 5.9 se aprecia que, también en esta condición, el criterio MAD es preciso únicamente para la clasificación de ítems con errores de especificación en los atributos difíciles (a diferencia de las condiciones anteriores es sensible también al error de dependencia secuencial simple). Las FVP oscilan entre 0.72 y 0.99 indicando una alta precisión en la clasificación de los ítems en las condiciones de error de especificación mencionadas. Los puntos de corte muestran que con valores MAD por encima de 0.043 se detecta con precisión la presencia

de errores de especificación en los atributos difíciles. El comportamiento de los valores LSD es muy similar al encontrado en las condiciones previas, siendo poco sensible a los errores de especificación de **Q**.

Tabla 5.9: Estadísticos ROC del criterio MAD en la condición 8A15I

Q	MAD* (CI 95 %)	<i>sd</i>	FVP (CI 95 %)	<i>sd</i>	PAUC (CI 95 %)
AD	0.038 (0.037-0.039)	0.016	0.094 (0.091-0.096)	0.037	0.057 (0.053-0.061)
DE	0.096 (0.092-0.099)	0.058	0.803 (0.792-0.813)	0.169	0.417 (0.401-0.433)
DOF	0.043 (0.041-0.044)	0.018	0.864 (0.854-0.873)	0.148	0.542 (0.527-0.556)
DI	0.050 (0.048-0.051)	0.025	0.707 (0.701-0.712)	0.085	0.438 (0.424-0.452)
DAF	0.073 (0.068-0.077)	0.073	0.769 (0.764-0.773)	0.064	0.443 (0.421-0.464)
DIN	0.047 (0.045-0.048)	0.022	0.867 (0.859-0.874)	0.121	0.539 (0.522-0.555)
DDS	0.050 (0.048-0.051)	0.025	0.730 (0.723-0.736)	0.097	0.442 (0.427-0.456)
FE	0.060 (0.057-0.062)	0.033	0.277 (0.272-0.281)	0.065	0.145 (0.138-0.153)
FOF	0.043 (0.041-0.044)	0.020	0.270 (0.264-0.275)	0.088	0.122 (0.115-0.129)
FI	0.075 (0.072-0.077)	0.040	0.326 (0.319-0.332)	0.100	0.164 (0.157-0.172)
FAF	0.051 (0.049-0.052)	0.023	0.130 (0.126-0.133)	0.058	0.076 (0.071-0.080)
FIN	0.056 (0.054-0.057)	0.027	0.399 (0.390-0.407)	0.134	0.190 (0.181-0.199)
FDS	0.056 (0.054-0.057)	0.028	0.375 (0.367-0.382)	0.126	0.197 (0.189-0.204)
DSC	0.016 (0.015-0.016)	0.006	0.375 (0.367-0.382)	0.126	0.197 (0.189-0.204)

* Valor promedio del estadístico MAD asociado a una FFP de 0.05

Tabla 5.10: Estadísticos ROC del criterio LSD en la condición 8A15I

Q	LSD (CI 95 %)	<i>sd</i>	FVP (CI 95 %)	<i>sd</i>	PAUC (CI 95 %)
AD	0.685 (0.681-0.688)	0.061	0.072 (0.068-0.075)	0.057	0.025 (0.021-0.029)
DE	0.880 (0.874-0.885)	0.086	0.207 (0.204-0.209)	0.044	0.164 (0.159-0.168)
DOF	0.931 (0.923-0.938)	0.112	0.212 (0.209-0.214)	0.038	0.170 (0.165-0.185)
DI	0.649 (0.644-0.653)	0.068	0.120 (0.111-0.128)	0.134	0.030 (0.025-0.035)
DAF	0.723 (0.718-0.727)	0.064	0.165 (0.157-0.172)	0.120	0.054 (0.048-0.059)
DIN	0.742 (0.737-0.746)	0.068	0.170 (0.162-0.177)	0.115	0.058 (0.052-0.064)
DDS	0.872 (0.866-0.877)	0.080	0.213 (0.209-0.216)	0.049	0.166 (0.161-0.172)
FE	0.686 (0.683-0.688)	0.045	0.098 (0.093-0.102)	0.074	0.028 (0.024-0.033)
FOF	0.785 (0.781-0.788)	0.056	0.134 (0.130-0.137)	0.061	0.075 (0.070-0.081)
FI	0.659 (0.654-0.663)	0.067	0.075 (0.070-0.079)	0.070	0.020 (0.016-0.023)
FAF	0.716 (0.712-0.719)	0.060	0.104 (0.099-0.108)	0.065	0.036 (0.031-0.041)
FIN	0.683 (0.679-0.686)	0.059	0.081 (0.077-0.084)	0.061	0.024 (0.020-0.027)
FDS	0.865 (0.860-0.869)	0.066	0.156 (0.152-0.159)	0.057	0.111 (0.105-0.116)
DSC	0.662 (0.656-0.667)	0.091	0.156 (0.152-0.159)	0.057	0.111 (0.105-0.116)

* Valor promedio del estadístico LSD asociado a una FFP de 0.05

En las Tablas 5.11 y 5.12 se resumen los resultados de los estadísticos ROC para la condición 8A32I. Al igual que en las condiciones previas se aprecia que el criterio MAD es sensible a errores de especificación que afectan a los atributos difíciles, incluidos los errores de dependencia secuencial simple y compleja. Los puntos de corte indican que

con valores MAD por encima de 0.053 se detectaría la presencia de este tipo de errores de especificación. A pesar de la alta tasa de verdaderos positivos evidenciada en esta condición (valores sobre 0.80), sólo dos valores del área parcial estandarizada (PAUC) superan el valor de 0.5, las restantes áreas se aproximan a 0.4, indicando baja precisión del clasificador en el intervalo de FFN entre 0 y 0.5.

Tabla 5.11: Estadísticos ROC del criterio MAD en la condición 8A32I

Q	MAD* (CI 95 %)	<i>sd</i>	FVP (CI 95 %)	<i>sd</i>	PAUC (CI 95 %)
AD	0.069 (0.067-0.070)	0.025	0.114 (0.112-0.115)	0.028	0.062 (0.059-0.064)
DE	0.057 (0.055-0.058)	0.026	0.800 (0.794-0.805)	0.094	0.380 (0.369-0.391)
DOF	0.053 (0.051-0.054)	0.021	0.842 (0.836-0.847)	0.081	0.514 (0.500-0.528)
DI	0.058 (0.056-0.059)	0.028	0.854 (0.852-0.855)	0.030	0.443 (0.433-0.453)
DAF	0.057 (0.055-0.058)	0.028	0.951 (0.947-0.954)	0.050	0.423 (0.410-0.436)
DIN	0.057 (0.055-0.058)	0.027	0.980 (0.977-0.982)	0.038	0.488 (0.473-0.503)
DDS	0.062 (0.059-0.064)	0.034	0.829 (0.824-0.833)	0.067	0.349 (0.340-0.358)
FE	0.062 (0.060-0.063)	0.026	0.406 (0.397-0.414)	0.129	0.158 (0.150-0.165)
FOF	0.066 (0.063-0.068)	0.034	0.285 (0.279-0.290)	0.088	0.118 (0.114-0.122)
FI	0.064 (0.062-0.065)	0.031	0.401 (0.395-0.406)	0.094	0.171 (0.167-0.176)
FAF	0.069 (0.067-0.070)	0.028	0.198 (0.193-0.202)	0.074	0.077 (0.073-0.080)
FIN	0.071 (0.069-0.072)	0.028	0.409 (0.400-0.417)	0.134	0.157 (0.152-0.162)
FDS	0.068 (0.065-0.070)	0.033	0.358 (0.351-0.364)	0.108	0.129 (0.123-0.134)
DSC	0.052 (0.050-0.053)	0.022	0.862 (0.858-0.865)	0.051	0.555 (0.547-0.563)

* Valor promedio del estadístico MAD asociado a una FFP de 0.05

Tabla 5.12: Estadísticos ROC del criterio LSD en la condición 8A32I

Q	LSD (CI 95 %)	<i>sd</i>	FVP (CI 95 %)	<i>sd</i>	PAUC (CI 95 %)
AD	0.550 (0.547-0.552)	0.040	0.114 (0.110-0.117)	0.048	0.058 (0.054-0.063)
DE	0.647 (0.641-0.652)	0.090	0.188 (0.185-0.190)	0.042	0.134 (0.129-0.138)
DOF	0.621 (0.616-0.625)	0.073	0.238 (0.235-0.240)	0.038	0.175 (0.170-0.180)
DI	0.483 (0.481-0.484)	0.021	0.245 (0.237-0.252)	0.113	0.066 (0.060-0.072)
DAF	0.528 (0.526-0.529)	0.026	0.259 (0.254-0.263)	0.074	0.103 (0.095-0.112)
DIN	0.574 (0.570-0.577)	0.048	0.290 (0.289-0.298)	0.058	0.180 (0.172-0.189)
DDS	0.530 (0.527-0.532)	0.035	0.218 (0.213-0.222)	0.073	0.442 (0.436-0.448)
FE	0.647 (0.642-0.651)	0.077	0.174 (0.171-0.176)	0.044	0.112 (0.107-0.116)
FOF	0.583 (0.579-0.586)	0.048	0.136 (0.132-0.139)	0.053	0.076 (0.072-0.080)
FI	0.511 (0.508-0.513)	0.032	0.141 (0.136-0.145)	0.067	0.051 (0.046-0.056)
FAF	0.523 (0.520-0.525)	0.042	0.114 (0.110-0.117)	0.058	0.049 (0.045-0.054)
FIN	0.488 (0.485-0.490)	0.038	0.119 (0.114-0.123)	0.070	0.039 (0.035-0.043)
FDS	0.550 (0.546-0.553)	0.054	0.106 (0.102-0.109)	0.049	0.056 (0.051-0.061)
DSC	0.616 (0.611-0.620)	0.064	0.463 (0.460-0.465)	0.042	0.394 (0.390-0.397)

* Valor promedio del estadístico LSD asociado a una FFP de 0.05

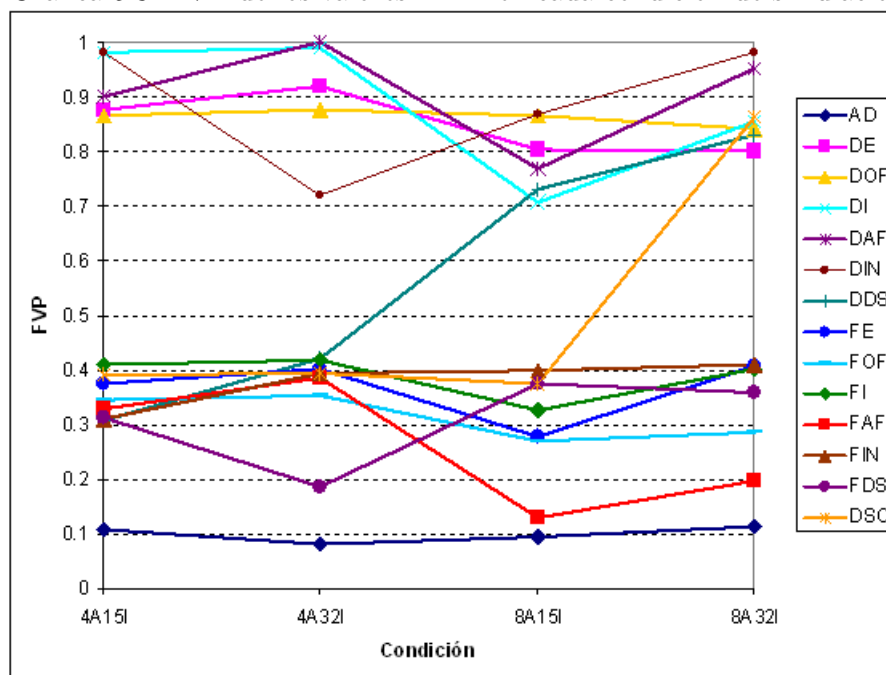
En cuanto al criterio LSD (véase Tabla 5.12) los resultados son muy similares a los encontrados en otras condiciones ya que dichos valores no presentan buena sensibilidad

para la detección de ningún tipo de error de especificación de las matrices \mathbf{Q} , esto se evidencia con valores medios de FVP inferiores a 0.46.

5.3.1. Comparación entre condiciones

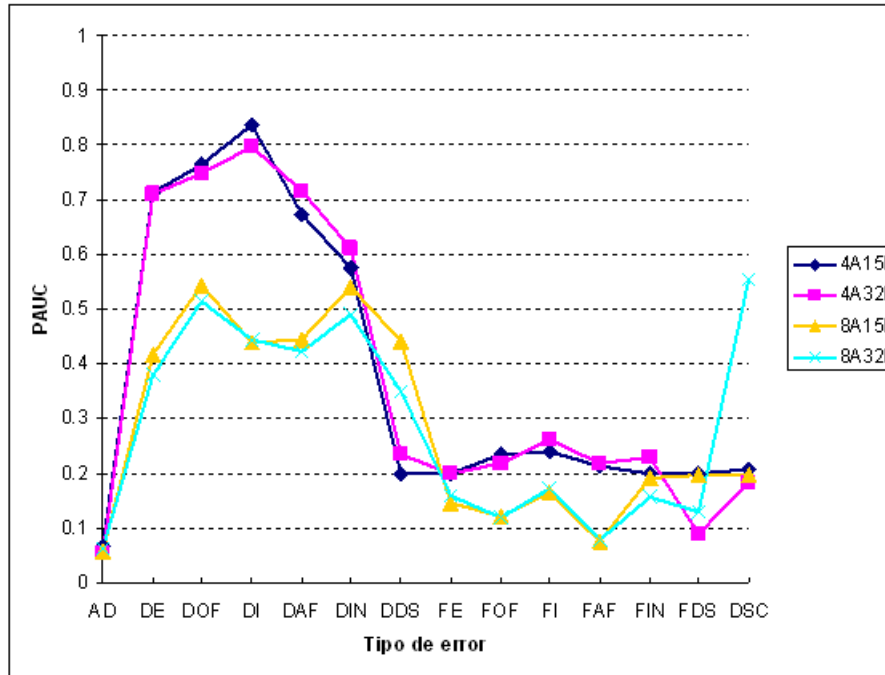
En la Gráfica 5.3 se representa la fracción de verdaderos positivos de los valores MAD de las cuatro condiciones en los 14 tipos de error introducidos en la matriz \mathbf{Q} . No se presentan los valores LSD porque no han sido sensibles a ninguna condición. En dicha gráfica se puede evidenciar que el MAD es altamente sensible a las condiciones DE, DOF, DI, DAF y DIN, es decir aquellas que afectan a los atributos difíciles independientemente de si se cambian ceros por unos (inclusión) o unos por ceros (exclusión). Es interesante notar que los errores de dependencia secuencial simple y compleja (DDS y DSC) son las únicas condiciones de error que cambian de acuerdo a las condiciones de diseño, ya que los valores MAD resultan sensibles a estos tipos de error en las condiciones con 8 atributos pero no cuando la estructura contiene 4 atributos. También se puede evidenciar que el estadístico MAD no permite clasificar con precisión a los ítems en las condiciones en las que se alteran los atributos fáciles, este resultado es homogéneo en las cuatro condiciones de diseño. También resulta notorio que el estadístico es muy poco sensible a la adición de atributos irrelevantes en la estructura cognitiva (AD).

Gráfica 5.3: FVP de los valores MAD en cada condición de simulación



En la Gráfica 5.4 se presentan los valores PAUC para el estadístico MAD, en la gráfica se puede ver que las áreas superan el valor 0.50 en las condiciones DE, DOF, DI, DAF y DIN cuando la matriz \mathbf{Q} tiene 4 atributos. Para las estructuras con 8 atributos las áreas parciales se encuentran por debajo de dicho valor indicando que la clasificación de los ítems es aleatoria y que, por tanto, el estadístico no detecta adecuadamente los errores de especificación de la matriz \mathbf{Q} , lo anterior no se cumple en la condición 8A32I cuya área parcial supera el valor de 0.5 en la condición de dependencia secuencial compleja.

Gráfica 5.4: PAUC de los valores MAD en cada condición de error



5.4. Conclusiones

En el presente estudio se han analizado las propiedades estadísticas de los criterios de validación propuestos por Dimitrov (2007) para el LSDM: los valores MAD, que cuantifican la calidad de la recuperación de las CCIs y los valores LSD, que informan sobre la precisión de la solución minimo-cuadrática encontrada. Lo anterior se ha realizado por medio del estudio de la sensibilidad y la especificidad de dichos criterios para la detección de varios tipos de error de especificación introducidos en las matrices \mathbf{Q} de cuatro condiciones. En general, los resultados muestran que el clasificador MAD presenta buena precisión (sensibilidad y especificidad) para la detección de errores de especificación introducidos en atributos con alta dificultad relativa. Por el contrario, dicho clasificador no presenta buenas propiedades para la detección de errores de especificación introducidos en atributos fáciles. Este resultado confirma lo encontrado en el Capítulo 4 de la presente tesis, en el cual se realizó un estudio descriptivo de los efectos de los errores de especificación, pues en dicho trabajo los errores de especificación en los atributos difíciles alteraban notablemente la estimación de la probabilidad de dominio de los atributos y la forma de las CPAs, pero no ocurría lo mismo cuando se modificaban los atributos fáciles (ver apartado 4.4 del

Capítulo 4).

En cuanto a los puntos de corte se ha encontrado cierta estabilidad a lo largo de los tipos de error introducidos en \mathbf{Q} y también en las cuatro condiciones experimentales. Valores MAD sobre 0.04 indican la presencia de errores de especificación en los atributos difíciles de la estructura establecida en \mathbf{Q} . En el análisis de la recuperación de las CCIs con el LSDM (Capítulo 3) se identificaron puntos de corte para el estadístico MAD en condiciones en las que se analizaban los datos con la TRI y se empleaban los parámetros del modelo de Rasch para ejecutar el método LSDM. Los puntos de corte oscilaban entre 0.02 y 0.12 (ver Tabla 3.7 del Capítulo 3) dependiendo principalmente del número de atributos requeridos por el ítem. En dicho caso el incremento de los valores MAD puede atribuirse al efecto de la precisión de la estimación de los parámetros en la TRI. En este caso, por el contrario, el incremento de dicho estadístico se debe únicamente a la presencia de errores de especificación por lo que a nivel aplicado hay que tener en cuenta que el valor MAD puede incluir una combinación de las dos fuentes de error, en otras palabras, un valor MAD que sobrepase el punto de corte para un ítem específico puede estar indicando una mala especificación de la matriz \mathbf{Q} pero también puede deberse a fluctuaciones atribuidas a la estimación de los parámetros en la TRI.

También es interesante notar que no existe mucha diferencia en el comportamiento de los estadísticos entre las diversas condiciones de la simulación, evidenciando que no hay efectos notorios del número de atributos y de ítems que componen el test. Un resultado que merece ser discutido es que las dependencias secuenciales no parecen ser detectados por el estadístico MAD en las condiciones de 4 atributos pero sí en las de 8, una posible explicación es que en el último caso hay mayor número de cambios en \mathbf{Q} y como hay más atributos involucrados también hay mayor complejidad en las relaciones de dependencia. Otro resultado importante es que el LSD no es sensible a ningún tipo de error de especificación de la matriz \mathbf{Q} ,

A pesar de la falta de sensibilidad de los criterios del LSDM para la detección de algunos tipos de error de especificación en la matriz \mathbf{Q} el criterio MAD ha resultado ser útil para la validación de ciertos tipos de estructuras cognitivas, específicamente para aquellas que involucran atributos difíciles, por lo que el método puede ser una buena opción para acopiar evidencia de validez en los procesos de diagnóstico cognitivo.

Capítulo 6

Aplicaciones del LSDM

6.1. Introducción

El conocimiento sobre estructuras cognitivas que subyacen al desempeño en ítems de los test puede ayudar a los desarrolladores de pruebas y a los educadores a construir ítems con determinadas características tanto a nivel de la exigencia cognitiva como a propiedades de medida deseables. Un entendimiento más profundo de los atributos de pensamiento, aprendizaje y habilidades involucradas en la solución de los test tiene muchas ventajas que incluyen un mejor diagnóstico de las fortalezas y dificultades de los examinados y el desarrollo de estrategias de enseñanza más enfocadas en procesos cognitivos específicos.

Por los anteriores motivos, el objetivo principal del presente estudio es analizar los atributos que subyacen a la ejecución de los estudiantes en un test de operaciones aritméticas básicas entre números enteros utilizando el LSDM (Dimitrov, 2007) y explorar la capacidad del método como herramienta de diagnóstico cognitivo. Para alcanzar dicha meta, el presente trabajo tiene los siguientes objetivos específicos:

1. Acopiar evidencia sobre la validez de la estructura cognitiva propuesta para un test de aritmética diseñado bajo una perspectiva psicométrico-cognitiva.
2. Evaluar la calidad de la recuperación de las CCIs por medio del análisis de los valores MAD, MAD ponderado (WMAD) y de la Raíz de la Diferencia Media al Cuadrado (RMSD). También se propone una nueva aproximación al estudio de la recuperación de las CCIs, basada en el trazo de límites superiores e inferiores a la curva.

3. Proponer una metodología para usar el LSDM como herramienta para el diagnóstico cognitivo a través de la clasificación de examinados en estados de conocimiento.
4. Utilizar los modelos LLTM (Fischer, 1973) y DINA (Junker y Sijtsma, 2001) para realizar una validación cruzada de los resultados obtenidos.
5. Analizar la exactitud del LSDM para clasificar a los examinados en estados de conocimiento mediante la comparación de los resultados de los modelos LSDM y DINA con los estados de conocimiento observados que corresponden a la proporción de acierto de ítems individuales diseñados para medir el desempeño de los sujetos en cada atributo.

6.2. Procedimiento

El presente trabajo se ha realizado en seis fases que se describirán a continuación:

6.2.1. Especificación de la estructura cognitiva

En primer lugar se propone una estructura cognitiva para explicar el proceso de solución de tareas de aritmética que incluyen operaciones de suma y resta con números enteros. Dicha estructura se ha basado en: a) entrevistas a profesores de Matemáticas, b) estudio de los contenidos de las asignaturas de Matemáticas en los colegios objetivo y c) teorías cognitivas en las que se ha realizado un análisis sistemático de los procesos y estrategias de solución de tareas de adición y sustracción con números enteros (Campbell, 2005; Real, Olea, Ponsoda, Revuelta y Abad, 1999; Kamii y Joseph, 1989; Maza, 1991, 2001; Mialaret, 1984). Una presentación detallada de dichas teorías y de las bases para la estructura cognitiva planteada para el presente estudio se encuentra en el Capítulo 2 (Marco Teórico). En dicho Capítulo se plantean los siguientes procesos para la solución de las tareas de aritmética con números enteros:

- a) Simplificación o resolución de paréntesis, que involucra dos tipos de reglas $(a - (-b) = a + b$ y $a + (-b) = a - b$).
- b) Clasificación de la operación. Después de la simplificación, las operaciones pueden ser de cinco clases $(a + b, a - b$ con $a > b, a - b$ con $a < b, -a + b$ y $-a - b)$ y según

la clase a la que pertenezca tienen un proceso de solución diferente e involucran el dominio de alguno de los atributos consecuentes.

- c) Adición entre números enteros y resta cuando el primer componente es menor que el segundo: corresponde a las operaciones del tipo $-a + b$ y $a - b$ con $a < b$, para resolverlas es necesario tomar el valor absoluto de los dos números, restar el menor del mayor y poner al resultado el signo del número mayor en valor absoluto.
- d) Sustracción entre números enteros: se refiere a operaciones del tipo $-a - b$, para resolverlas es necesario tomar el valor absoluto de los dos números, sumarlos y asignar signo negativo al resultado.
- e) Suma y resta simples: corresponden a operaciones entre naturales de tipo $a + b$ y $a - b$ con $a > b$.

Se tiene la hipótesis de que los dos últimos atributos son básicos y necesarios para el dominio de los demás atributos, son los que primero se enseñan y se considera que serán los más fáciles para los examinados pues han sido aprendidos en cursos anteriores a los correspondientes a la aplicación del test (que ha sido diseñado para estudiantes de séptimo grado de secundaria). Como se trata de dos procesos que ya están automatizados en los estudiantes de dicha edad y nivel educativo, se tiene la hipótesis de que los sujetos utilizan estrategias de *recuperación de la memoria* para resolverlos. Dicha estrategia no implica una descomposición en procesos ni la aplicación de reglas sino una recuperación directa de la solución (Campbell, 2005).

Los atributos cognitivos anteriormente descritos se han numerado y ordenado como se expone en la Tabla 6.1. Nótese que se ha prescindido del proceso de clasificación ya que es necesario en todas las tareas.

Tabla 6.1: Estructura cognitiva del Test de Aritmética

Atributo	Símbolo	Descripción
A_1 : suma	$a + b = c$	Adiciona las cantidades
A_2 : resta	$a - b$ $a > b$	Sustraer las cantidades
A_3 : operación enteros 1	$-a + b = b - a$ $a - b$ $a < b$	Toma el valor absoluto, resta el número mayor del menor, pone al resultado el signo del componente mayor
A_4 : operación enteros 2	$-a - b = -(a + b)$	Toma el valor absoluto, suma los números y pone signo negativo al resultado.
A_5 : Simplificación 1	$a - (-b) = a + b$ $-a - (-b) = -a + b$	Cambia el signo de b y cambia el operador resta por suma.
A_6 : Simplificación 2	$a + (-b) = a - b$ $-a + (-b) = -a - b$	Cambia el signo de b y cambia el operador suma por resta.

6.2.2. Diseño y aplicación del Test de Aritmética

En esta fase se ha diseñado un Test de Aritmética (TA) de acuerdo a la estructura cognitiva anteriormente descrita. El TA (presentado en el Anexo 3 del Capítulo 8) se compone de 39 ítems de selección múltiple, divididos en dos grupos:

1. Items Individuales: el primer grupo lo componen 26 ítems individuales que se han diseñado para medir el desempeño de los examinados en cada uno de los atributos. Por medio de dichos ítems se obtienen puntuaciones observadas de la ejecución de los estudiantes en cada una de las seis operaciones propuestas. Un ejemplo de este tipo de ítems es: **-9-(-5)** = a) -9+5; b) -9-5; c) 9+5; d) 9-5 (véase Anexo 3). Como se puede observar el ítem no pregunta directamente al examinado por la respuesta total de la operación sino que busca evaluar el dominio específico del atributo A_5 , en otras palabras, por medio de dicho ítem se pretende conocer si el estudiante sabe que para resolver un paréntesis de este tipo hay que cambiar el signo del componente b y el operador resta por suma. Las opciones incorrectas constituyen aplicaciones erróneas de dicha regla, por ejemplo, en la opción b) se cambia el signo del componente b pero no el del operador (resta por suma), en la opción c) se cambia erróneamente el signo del componente a y en la opción d) no se cambia el operador y se cambia el signo del componente a .
2. Items Compuestos: el segundo grupo se compone de 13 ítems que requieren el dominio de varios atributos para ser resueltos correctamente. En este tipo de tarea se le pide al alumno la solución total de la operación y no los pasos para resolverla. Por ejemplo, para resolver correctamente el ítem: **-2+(-8)** = a) -10; b) -6; c) 6; d) 10 (véase Anexo 3) se debe aplicar primero el atributo de solución de paréntesis A_6 , esto lleva a la expresión $-2 - 8$, posteriormente se deberá utilizar el atributo A_4 , es decir se suman las cantidades y se antepone signo negativo al resultado, conduciendo a la operación $-(2 + 8)$. Finalmente, para sumar las cantidades se requiere también el dominio de la suma simple (A_1) con lo que se llega al resultado correcto -10 . Los ítems compuestos se generaron de acuerdo a los 10 posibles tipos de tareas que se pueden conformar manipulando el signo de los componentes, su valor absoluto y los operadores (suma o resta):

- a) Dos ítems (i3 e i19) del tipo $-a + (-b)$ que requieren, en su orden, los procesos A_5 , A_4 y A_1
- b) Un ítem (i18) del tipo $-a - b$ que requiere, en su orden, los procesos A_4 y A_1
- c) Un ítem (i4) del tipo $-a - (-b)$ con $a > b$ que requiere, en su orden, los procesos A_6 y A_3
- d) Un ítem (i24) del tipo $-a - (-b)$ con $a \leq b$ que requiere, en su orden, los procesos A_6 , A_3 y A_2
- e) Un ítem (i10) del tipo $-a + b$ con $a > b$ que requiere el proceso A_3
- f) Dos ítems (i21 e i31) del tipo $-a + b$ con $a \leq b$ que requieren, en su orden, los procesos A_3 y A_2
- g) Un ítem (i35) del tipo $a + (-b)$ con $a > b$ que requiere en su orden los procesos A_5 y A_2
- h) Dos ítems (i16 e i29) del tipo $a + (-b)$ con $a \leq b$ que requiere en su orden los procesos A_5 y A_3
- i) Un ítem (i14) del tipo $a - (-b)$ que requiere, en su orden, los procesos A_6 y A_1
- j) Un ítem (i23) del tipo $a - b$ con $a < b$ que requiere el proceso A_3

Los vínculos entre los 6 atributos y los 39 ítems del TA se expresan en la matriz **Q** presentada en la Tabla 6.2. En la última columna de dicha tabla se incluye el número total de atributos requeridos por cada ítem compuesto y, en caso de tratarse de un ítem individual, se presenta el atributo que pretende evaluar.

Tabla 6.2: Estructura cognitiva del TA (matriz **Q**)

	Item	A_1	A_2	A_3	A_4	A_5	A_6	Total
1	2+8	1	0	0	0	0	0	A1
2	-9-(-5)	0	0	0	0	1	0	A5
3	-2+(-8)	1	0	0	1	1	0	3
4	-9-(-5)	0	0	1	0	0	1	2
5	7-5	0	1	0	0	0	0	A2
6	-9+(-7)	0	0	0	0	0	1	A6
7	9+7	1	0	0	0	0	0	A1
8	3-7	0	0	1	0	0	0	A3
9	8+(-4)	0	0	0	0	0	1	A6
10	-6+3	0	0	1	0	0	0	1
11	-5+7	0	0	1	0	0	0	A3
12	-6-9	0	0	0	1	0	0	A4
13	-7-(-9)	0	0	0	0	1	0	A5
14	2-(-7)	1	0	0	0	0	1	2
15	-9+9	0	0	1	0	0	0	A3
16	0+(-6)	0	0	1	0	1	0	2
17	2-(-7)	0	0	0	0	1	0	A5
18	-6-9	1	0	0	1	0	0	2
19	-9+(-7)	1	0	0	1	1	0	3
20	0+(-6)	0	0	0	0	0	1	A6
21	-9+9	0	1	1	0	0	0	2
22	3+(-7)	0	0	0	0	0	1	A6
23	5-7	0	0	1	0	0	0	1
24	-7-(-9)	0	1	1	0	0	1	3
25	-6+3	0	0	1	0	0	0	A3
26	-2+(-8)	0	0	0	0	0	1	A6
27	5-9	0	0	1	0	0	0	A3
28	-9+5	0	0	1	0	0	0	A3
29	3+(-7)	0	0	1	0	1	0	2
30	-7+9	0	0	1	0	0	0	A3
31	-5+7	0	1	1	0	0	0	2
32	-9-7	0	0	0	1	0	0	A4
33	2-0	0	1	0	0	0	0	A2
34	-6-0	0	0	0	1	0	0	A4
35	8+(-4)	0	1	0	0	1	0	2
36	8-4	0	1	0	0	0	0	A2
37	3-6	0	0	1	0	0	0	A3
38	-2-8	0	0	0	1	0	0	A4
39	2+7	1	0	0	0	0	0	A1

Cada uno de los pasos (o componentes) requeridos para solucionar los ítems compuestos fueron medidos a través de los ítems individuales, sin embargo cada ítem compuesto y sus respectivos individuales no se presentan de forma continua en la prueba sino que su posición ha sido asignada aleatoriamente. Otra de las variables que se ha controlado para el diseño

de la prueba es el tamaño de los componentes: se han incluido componentes menores a 10 puesto que los algoritmos y estrategias de solución para tareas con valores mayores a 10 son diferentes de los aquí contemplados.

El test de aritmética ha sido aplicado a la muestra de 382 estudiantes de 7 grado de secundaria en cinco colegios públicos de Bogotá-Colombia. Los alumnos han contado con una hora, durante la clase de Matemáticas, para resolver la prueba, se ha explicado a los alumnos que la prueba hacía parte de los ejercicios del tema que se encontraban estudiando en ese momento y que, aunque su calificación no se incluiría en la nota general de la asignatura, sí daría puntos al igual que un ejercicio de clase.

6.2.3. Análisis mediante la TRI

En esta fase se ha analizado la dimensionalidad de los datos mediante la aplicación del modelo de ojiva normal (NOHARM, Fraser y McDonald, 1988), para valorar el grado de unidimensionalidad se ha usado el índice de Tanaka (Tanaka y Huba, 1985), se han analizado la raíz media de los residuales al cuadrado (RMSR) y las saturaciones de los ítems al factor. Posteriormente se ha analizado la matriz de datos con el modelo TRI y se ha evaluado su ajuste.

6.2.4. Aplicación del método LSDM

En este paso se ha aplicado el método LSDM al conjunto de ítems y se han ejecutado varios procedimientos para recabar evidencia sobre la validez de la estructura cognitiva: se ha evaluado la adecuación de los vínculos entre ítems y atributos mediante el análisis de la recuperación de las CCIs por medio de los estadísticos MAD, WMAD y RMSD. También se han analizado los residuales (LSD) y el comportamiento de las CPAs. El método LSDM y los estadísticos de recuperación se han ejecutado mediante una rutina programada en el entorno R.

6.2.5. Análisis de los atributos según su dificultad

En esta fase se han aplicado a los datos del TA los modelos LLTM y DINA, se ha analizado su ajuste y se han comparado los resultados de los tres modelos (LLTM, DINA y LSDM) en cuanto al orden de la dificultad relativa de los atributos. Para el análisis con

el LLTM se ha empleado el programa LPCM-WIN (Fischer y Ponocny-Seliger, 1998) y para la estimación del modelo DINA se ha usado una rutina escrita en Ox (de la Torre, 2008), que ha sido requerida a su autor. La estimación de parámetros con el modelo DINA se ha llevado a cabo por máxima verosimilitud marginal usando el algoritmo EM.

6.2.6. Clasificación en estados de conocimiento

En esta fase se han obtenido tres matrices con las probabilidades de dominio de los atributos para cada uno de los 382 estudiantes: una con el método LSDM, otra con el modelo DINA y la tercera se ha denominado matriz empírica y corresponde a la proporción de acierto de cada sujeto a los ítems individuales. Finalmente se compara la clasificación obtenida con el método LSDM y el modelo DINA con la proporción de acierto observada en los ítems individuales.

6.3. Métodos para evaluar la recuperación de las CCIs

En el presente estudio se han empleado cuatro métodos para estudiar la recuperación de las CCIs con el LSDM. Además de utilizar el estadístico MAD (Dimitrov, 2007) se proponen tres métodos nuevos para evaluar la calidad de la recuperación de las CCIs:

6.3.1. Diferencia Media Absoluta

$$MAD_j = \frac{\sum_{m=1}^M |P_{jm} - \hat{P}_{jm}|}{M} \quad (6.1)$$

Donde,

M es el número de niveles de habilidad (θ_i) previamente fijados.

P_{jm} es la probabilidad para el ítem j en el nivel de habilidad m obtenida con la TRI.

\hat{P}_{jm} es la probabilidad para el ítem j en el nivel de habilidad m , estimada por el LSDM

6.3.2. Diferencia Media Absoluta Ponderada

Los valores MAD son un promedio de las diferencias de las CCIs a lo largo de los niveles fijos de habilidad. Sin embargo, en los estudios realizados sobre este método (Dimitrov, 2007; Romero, Ordóñez, López y Navarro, 2009; Dimitrov, Romero, Ponsoda y Ximénez

2006, 2007) se observa que la recuperación de la CCI suele ser peor en los niveles bajos de habilidad ya que la solución del vector \mathbf{X} es menos precisa en dichos niveles, lo que se refleja en valores LSD altos. Con base a estos resultados en el presente estudio se propone un valor MAD ponderado (WMAD) en el que las diferencias de los niveles bajos, donde hay mayor imprecisión, tienen un menor peso que en los demás niveles de habilidad como se puede observar en la siguiente ecuación:

$$WMAD_j = \frac{\sum_{m=1}^M W_m |P_{jm} - \hat{P}_{jm}|}{\sum W_m} \quad (6.2)$$

Donde,

$$W_m = 1 - LSD$$

M es el número de niveles de habilidad (θ_i) previamente fijados.

P_{jm} es la probabilidad para el ítem j en el nivel de habilidad m obtenida con la TRI.

\hat{P}_{jm} es la probabilidad para el ítem j en el nivel de habilidad m , estimada por el LSDM.

6.3.3. Raíz Media de la Diferencia al Cuadrado

$$RMSE = \sqrt{\frac{\sum_{m=1}^M (P_{jm} - \hat{P}_{jm})^2}{M}} \quad (6.3)$$

Donde la notación es la misma que en la expresión anterior. Para interpretar los estadísticos se han utilizado puntos de corte hallados en los estudios de simulación (Capítulos 3 y 5) de la presente tesis.

6.3.4. Método de límites.

Se propone una metodología complementaria para evaluar la calidad de la recuperación de las CCIs por medio de la cual se busca controlar el efecto del error estándar de estimación de los parámetros de la TRI en la recuperación de las CCIs por medio del LSDM. Según este procedimiento para cada CCI se calcula un límite superior y uno inferior adicionando y sustrayendo el error estándar de estimación a su respectivo parámetro en la ecuación del modelo TRI como se puede observar en la siguiente expresión (para el

modelo TRI de Rasch que es el utilizado en el presente trabajo):

$$limsup = \frac{\exp[\theta_i - (b_j + 2SE_{b_j})]}{1 + \exp[\theta_i - (b_j + 2SE_{b_j})]} \quad (6.4)$$

$$liminf = \frac{\exp[\theta_i - (b_j - 2SE_{b_j})]}{1 + \exp[\theta_i - (b_j - 2SE_{b_j})]} \quad (6.5)$$

Donde,

$limsup$ y $liminf$ son los límites de la ICC y corresponden a la probabilidad superior e inferior del ítem j en los niveles fijos de habilidad θ_i .

b_j es el parámetro de dificultad del ítem j .

SE_{b_j} es el error estándar de estimación del parámetro de dificultad del ítem j .

Una vez constituidos los límites se pueden usar para valorar si la CCI recuperada por el LSDM se encuentra dentro de los límites o por el contrario, los excede y en que niveles de habilidad ocurre esto último. Este procedimiento es complementario a los estadísticos MAD, WMAD y RMSD en el estudio de la calidad de los vínculos entre ítems y atributos.

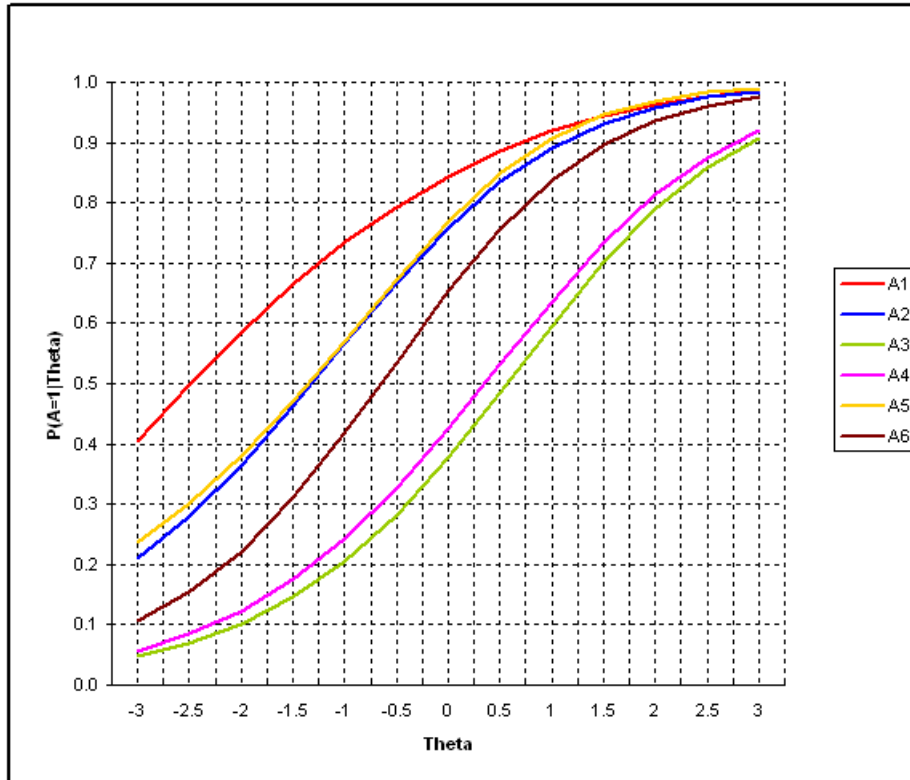
6.4. El LSDM como herramienta para el diagnóstico cognitivo

El método que se propone para utilizar el LSDM con fines diagnósticos es el siguiente:

1. Se cuenta con la habilidad estimada, para cada examinado, con el modelo de TRI.
2. Se obtienen, con el LSDM, las estimaciones de las probabilidades de dominio de cada atributo en los niveles fijos de habilidad, en este caso se han ampliado de 13 a 120 puntos de cuadratura para tener una estimación más precisa.
3. Se asigna a cada examinado una probabilidad de dominio de cada atributo de acuerdo a las CPAs estimadas por el LSDM. Por ejemplo, un sujeto cuyo nivel de habilidad estimado ha sido -2.5 tendrá una probabilidad de dominio de 0.5 para el atributo de suma A_1 , mientras que si su parámetro estimado es $\theta = 0.5$ se le asignará una probabilidad de dominio de 0.9 para dicha habilidad (véase Figura 6.1, dicha figura contiene las CPAs estimadas por el LSDM que serán analizadas en el apartado de

Resultados). Se procede de ésta manera para uno de los 382 estudiantes y para cada atributo obteniendo así una matriz de *estados de conocimiento* de 382 filas y 6 columnas que contiene las probabilidades de dominio de los atributos para cada uno de los alumnos examinados, de acuerdo a su nivel de habilidad estimado con la TRI.

Gráfica 6.1: Asignación de probabilidades de dominio de los atributos



6.5. Comparación de probabilidades de dominio

Una vez se ha obtenido la matriz de probabilidades de dominio usando el LSDM mediante el método propuesto anteriormente se estiman también en el modelo DINA. Finalmente se halla la matriz observada de dicha probabilidad que corresponde al cálculo, para cada examinado, de la proporción de acierto de los ítems individuales correspondientes a cada atributo.

El procedimiento final es comparar, para cada atributo, las matrices de probabilidad de dominio de los atributos obtenidas con el modelo DINA y LSDM con respecto a la matriz observada. Los estadísticos de comparación que serán presentados son la diferencia media

absoluta (MAD_k) y la raíz de la diferencia media al cuadrado ($RMSD_k$) expresadas en las ecuaciones 6.6 y 6.7 respectivamente:

$$MAD_k = \frac{\sum_{i=1}^I |P_i - \hat{P}_i|}{N} \quad (6.6)$$

Donde,

P_i es la proporción media de acierto para el examinado i en los ítems individuales que miden el desempeño en el atributo k .

\hat{P}_i es la probabilidad de dominio del atributo k para el examinado i estimada con el modelo DINA o con el LSDM.

$$N = 382$$

$$RMSD_k = \sqrt{\frac{\sum_{i=1}^I (P_i - \hat{P}_i)^2}{N}} \quad (6.7)$$

Donde la notación es la misma que en la expresión anterior.

6.6. Resultados

6.6.1. Análisis según la Teoría Clásica de los Test

La Tabla 6.3 muestra los estadísticos descriptivos y de la TCT para los ítems del TA. En la tabla se evidencian valores adecuados para todos los ítems, excepto la correlación biserial del ítem 24 con el test total, evidenciando que no es un ítem adecuado. La fiabilidad del test mediante el coeficiente alpha de Cronbach ha sido 0.791 indicando buena precisión en la medida de la habilidad aritmética. El coeficiente de correlación no se incrementa significativamente al excluir ninguno de los ítems. Las proporciones de acierto muestran que la prueba ha resultado difícil para los examinados. También se observa que los ítems más fáciles son los de suma (A_1) con una proporción de acierto media de 0.88, seguidos de los de resta cuando $a > b$ (A_2) con una proporción media de 0.59. Los ítems correspondientes a atributos de resolución de paréntesis tienen una proporción de acierto de 0.44 (A_5) y 0.31 (A_6), resultando más difícil la regla $a - (-b) = a + b$ que la regla $a + (-b) = a - b$. Los ítems individuales de mayor dificultad corresponden a los atributos A_3 y A_4 con una proporción media de acierto de 0.17 y 0.18 respectivamente. Este ordenamiento era el que

se esperaba de acuerdo a la teoría cognitiva pues los atributos A_3 y A_4 incluyen procesos complejos como tomar el valor absoluto y evaluar el signo del resultado, mientras que los atributos de suma y resta entre naturales son los que primero se aprenden y resultan básicos para las operaciones con números enteros.

Tabla 6.3: Estadísticos de la TCT para los ítems del TA

Item	Proporción		logit	ρ biserial		α de Cronbach
	0	1		incluido	excluido	excluido
1	0.105	0.895	2.146	0.257	0.206	0.791
2	0.709	0.291	-0.893	0.529	0.468	0.781
3	0.484	0.516	0.063	0.303	0.220	0.791
4	0.618	0.382	-0.480	0.220	0.137	0.794
5	0.380	0.620	0.491	0.375	0.299	0.788
6	0.662	0.338	-0.674	0.481	0.413	0.783
7	0.113	0.887	2.065	0.259	0.206	0.791
8	0.738	0.262	-1.037	0.267	0.194	0.792
9	0.476	0.524	0.094	0.307	0.225	0.791
10	0.668	0.332	-0.697	0.392	0.319	0.787
11	0.872	0.128	-1.916	0.314	0.260	0.789
12	0.822	0.178	-1.530	0.318	0.256	0.789
13	0.696	0.304	-0.830	0.521	0.459	0.781
14	0.728	0.272	-0.983	0.387	0.318	0.787
15	0.736	0.264	-1.023	0.228	0.153	0.793
16	0.772	0.228	-1.221	0.437	0.375	0.785
17	0.647	0.353	-0.604	0.414	0.341	0.786
18	0.374	0.626	0.514	0.286	0.205	0.791
19	0.479	0.521	0.084	0.287	0.204	0.792
20	0.529	0.471	-0.115	0.318	0.237	0.790
21	0.730	0.270	-0.996	0.421	0.354	0.786
22	0.482	0.518	0.073	0.356	0.277	0.789
23	0.581	0.419	-0.328	0.313	0.232	0.790
24	0.801	0.199	-1.393	0.089	0.019	0.797
25	0.866	0.134	-1.870	0.307	0.251	0.790
26	0.607	0.393	-0.436	0.337	0.259	0.789
27	0.767	0.233	-1.192	0.233	0.161	0.793
28	0.856	0.144	-1.783	0.260	0.202	0.791
29	0.683	0.317	-0.769	0.485	0.419	0.783
30	0.853	0.147	-1.762	0.274	0.216	0.791
31	0.764	0.236	-1.177	0.437	0.374	0.785
32	0.806	0.194	-1.426	0.246	0.180	0.792
33	0.458	0.542	0.168	0.374	0.295	0.788
34	0.840	0.160	-1.661	0.320	0.261	0.789
35	0.738	0.262	-1.037	0.509	0.448	0.782
36	0.387	0.613	0.458	0.407	0.332	0.786
37	0.812	0.188	-1.460	0.198	0.131	0.793
38	0.838	0.162	-1.641	0.286	0.226	0.790
39	0.149	0.851	1.741	0.275	0.216	0.791

6.6.2. Aplicación del modelo de TRI

6.6.2.1. Dimensionalidad y aplicación del modelo de Rasch

En la Tabla 6.4 se presentan los resultados del análisis de dimensionalidad y de TRI para los 39 ítems que componen el TA. El análisis de dimensionalidad presenta un índice Tanaka de 0.75 para el modelo unidimensional, por debajo del valor deseable (0.9; Tanaka y Huba, 1985). A pesar de que el valor del índice Tanaka es bajo, los residuales del modelo no son especialmente elevados con un valor RMSR de 0.02. Con una solución de dos dimensiones el índice Tanaka se eleva a 0.89 con un RMSR de 0.01. El análisis de las saturaciones indica que los ítems individuales conforman una dimensión y los ítems compuestos conforman otra. Por lo tanto se realizó de nuevo un análisis de la dimensionalidad separando los ítems individuales de los compuestos obteniéndose un índice Tanaka de 0.96 para el modelo unidimensional de los ítems compuestos ($\text{RMSR} = 0.02$) y de 0.89 para dicho modelo con los ítems individuales ($\text{RMSR} = 0.01$).

A pesar de que el TA es bidimensional se ha decidido analizar el test en conjunto pues las saturaciones de los ítems al factor son mayores a 0.20 excepto para los ítems número 24, 27 y 37. En la Tabla 6.4 se incluyen también las estimaciones de los parámetros de dificultad (b), con sus respectivos errores estándar (SE) y los estadísticos de ajuste para cada ítem. En general, los ítems presentan buen ajuste al modelo de Rasch con valores χ^2 pequeños, aunque la probabilidad del ítem 24 es inferior a 0.01 indicando desajuste, por dicho motivo y por sus malas propiedades psicométricas identificadas tanto en el análisis con la TCT como con la TRI se ha decidido eliminar dicho ítem para los análisis posteriores con el LSDM.

Tabla 6.4: Dimensionalidad y ajuste al modelo de Rasch para los 39 ítems del TA

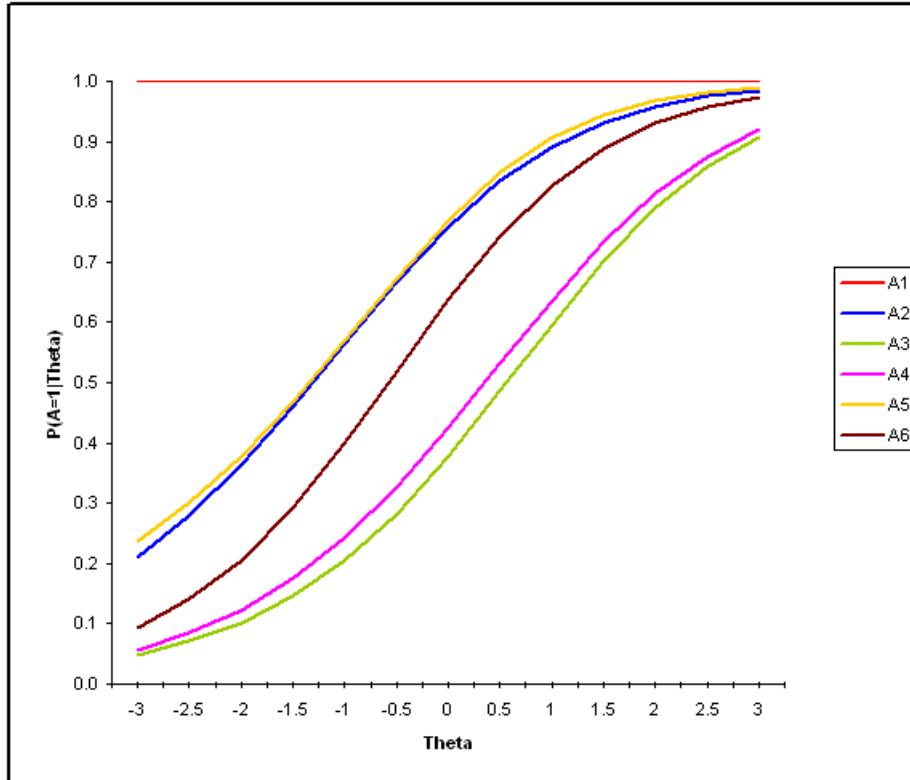
Cargas factor				Rasch			
		Item		b	SE	χ^2	$p(\chi^2)$
1	A1	2+8	0.426	-2.96	0.17	20.62	0.36
2	A6	-9-(-5)	0.726	0.33	0.12	33.11	0.02
3	G1	-2+(-8)	0.236	-0.74	0.11	24.57	0.18
4	G3	-9-(-5)	0.205	-0.13	0.11	28.33	0.08
5	A2	7-5	0.448	-1.21	0.11	20.01	0.39
6	A5	- 9+(-7)	0.603	0.09	0.12	26.41	0.12
7	A1	9+7	0.431	-2.88	0.16	15.62	0.68
8	A3	3-7	0.275	0.49	0.12	20.98	0.34
9	A5	8+(-4)	0.362	-0.77	0.11	31.01	0.03
10	G5	-6+3	0.482	0.11	0.12	15.70	0.68
11	A3	-5+7	0.405	1.47	0.16	24.17	0.19
12	A4	-6-9	0.414	1.04	0.14	15.92	0.66
13	A6	-7-(-9)	0.733	0.26	0.12	30.99	0.04
14	G8	2-(-7)	0.562	0.43	0.12	23.03	0.24
15	A3	-9+9	0.251	0.48	0.12	35.00	0.01
16	G6	0+(-6)	0.502	0.70	0.13	20.15	0.39
17	A6	2-(-7)	0.557	0.01	0.12	25.32	0.15
18	G2	-6-9	0.282	-1.23	0.11	22.14	0.28
19	G1	-9+(-7)	0.224	-0.76	0.11	20.44	0.37
20	A5	0+(- 6)	0.374	-0.54	0.11	21.91	0.29
21	G4	-9+9	0.485	0.45	0.12	14.97	0.72
22	A5	3+(-7)	0.439	-0.75	0.11	20.29	0.38
23	G5	5-7	0.338	-0.30	0.11	30.20	0.05
24	G9	-7-(-9)	0.060*	0.89	0.14	59.87	0.00*
25	A3	-6+3	0.351	1.42	0.16	20.06	0.39
26	A5	-2+(-8)	0.407	-0.18	0.11	12.64	0.86
27	A3	5-9	0.180	0.67	0.13	31.57	0.03
28	A3	-9+5	0.285	1.32	0.15	31.60	0.03
29	G6	3+(-7)	0.600	0.19	0.12	29.07	0.06
30	A3	-7+9	0.287	1.30	0.15	21.40	0.32
31	G4	-5+7	0.545	0.65	0.13	24.97	0.16
32	A4	-9-7	0.262	0.93	0.14	19.71	0.41
33	A2	2-0	0.407	-0.85	0.11	31.95	0.03
34	A4	-6-0	0.411	1.18	0.15	15.43	0.69
35	G7	8+(-4)	0.666	0.49	0.12	30.07	0.05
36	A2	8-4	0.488	-1.17	0.11	23.22	0.23
37	A3	3-6	0.147	0.97	0.14	20.66	0.36
38	A4	-2-8	0.329	1.16	0.15	30.43	0.05
39	A1	2+7	0.396	-2.54	0.15	16.00	0.66

6.6.3. Aplicación del LSDM

6.6.3.1. CPAs antes de la reespecificación:

En la Gráfica 6.2 se presentan las Curvas de Probabilidad de los Atributos (CPAs) obtenidas con el LSDM.

Gráfica 6.2: Curvas de probabilidad de los atributos sin reespecificar \mathbf{Q}



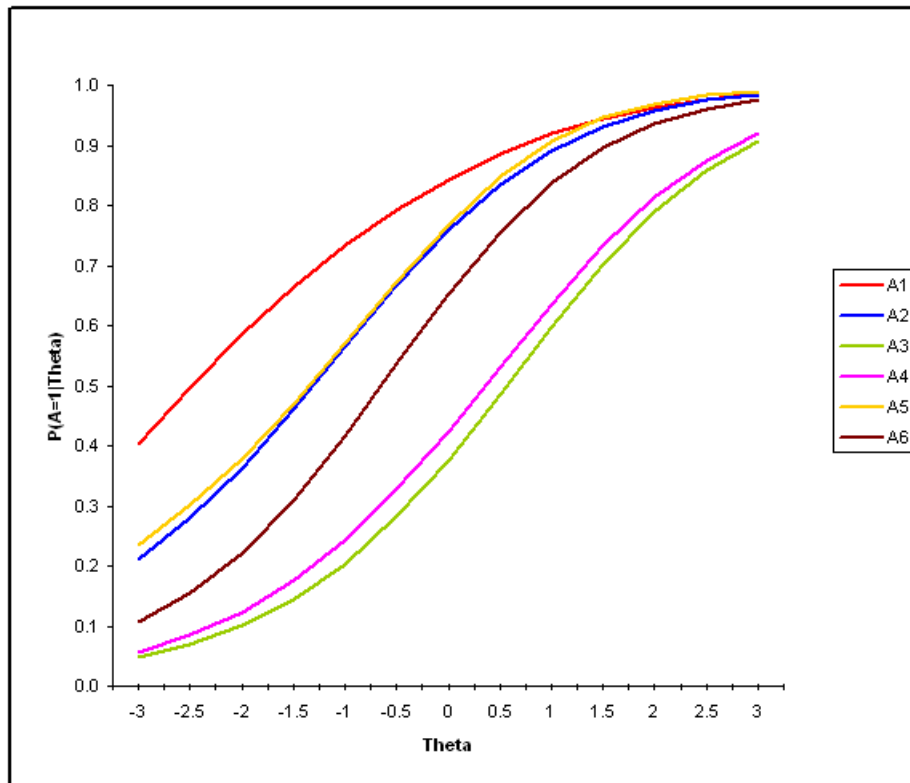
En la Gráfica 6.2 se puede observar que las probabilidades correspondientes al atributo de suma entre naturales (A_1) son una constante de 1 para todos los niveles de rasgo. Según los estudios de simulación de los capítulos precedentes este resultado indicaría que el dominio de dicha operación es irrelevante para la solución de este tipo de tareas, o que la habilidad representada en este atributo está contenida en otro componente, en particular en A_4 , pues hay una dependencia secuencial entre los dos atributos. En otras palabras, en los ítems compuestos, siempre que se utiliza A_1 está precedido por A_4 . Por el motivo anteriormente expuesto se ha decidido reespecificar la matriz \mathbf{Q} para poder obtener una estimación del atributo A_1 . Para llevarlo a cabo se ha eliminado el requerimiento de dicho atributo de los ítems compuestos que también necesitan el dominio de A_4 que son los ítems

3, 18 y 19; de esta manera se ha buscado eliminar la dependencia secuencial existente entre los dos atributos. Los valores LSD para esta estimación se encuentran entre 0.221 ($\theta = -3$) y 0.014 ($\theta = 3$).

6.6.3.2. CPAs después de la reespecificación:

En la Gráfica 6.3 se dibujan las CPAs obtenidas tras la reespecificación de la matriz **Q**. Como se esperaba se ha podido obtener una estimación del atributo A_1 pues se ha eliminado la relación de dependencia que tenía con el atributo A_4 . Adicionalmente, los valores LSD han disminuído tras la reespecificación encontrándose entre 0.186 ($\theta = -3$) y 0.008 ($\theta = 3$), dicho resultado también apoya la plausibilidad de la reespecificación realizada.

Gráfica 6.3: Curvas de probabilidad de los atributos tras la reespecificación de **Q**



En la Gráfica 6.3 se puede observar que el comportamiento de las curvas de probabilidad de los atributos es el esperado pues son monotónicas y el orden de la dificultad relativa de los atributos coincide con los lineamientos teóricos.

El atributo mas fácil ha sido el de suma, seguido por la resta cuando $a > b$ (A_2), y la

regla de simplificación $A_5 [a + (-b) = a - b]$. El atributo de dificultad media ha sido A_6 que corresponde a la otra regla de simplificación $[a - (-b) = a + b]$. Finalmente, como era de esperarse, los atributos más difíciles han sido A_3 y A_4 que involucran operaciones complejas como tomar el valor absoluto y valorar el signo del resultado. Como el comportamiento de las CPAs derivadas del análisis se corresponde adecuadamente con las hipótesis teóricas se procederá al análisis de recuperación de las CCIs.

Recuperación de las CCIs: En la Tabla 6.5 se presentan los estadísticos de recuperación de las CCIs (MAD, WMAD y RMSD) y la matriz \mathbf{Q} , ya reespecificada, para los 38 ítems del TA utilizados finalmente para el análisis. Para interpretar los estadísticos se han utilizado puntos de corte hallados en los estudios de simulación (Capítulos 3 y 5) de la presente tesis. En el estudio presentado en el Capítulo 3, en el que se utiliza la TRI (al igual que en el presente estudio) se ha encontrado que el punto de corte apropiado para los valores MAD en un escenario similar al que nos ocupa (6 atributos) es de 0.086 si el ítem requiere un atributo, 0.068 si requiere dos, 0.051 si requiere tres, 0.031 si requiere cuatro y 0.018 si requiere cinco o seis (ver Tabla 3.7 del Capítulo 3). Adicionalmente, valores MAD sobre 0.04 indican la presencia de errores de especificación en los atributos difíciles del ítem en cuestión (véase apartado 5.4 del Capítulo 5). Siguiendo dichos criterios, (véase Tabla 6.5) la CCI de 18 de los 39 ítems está mal recuperada mediante el análisis de los valores MAD. Es importante tener en cuenta que los puntos de corte obtenidos en el Capítulo 3 han sido obtenidos utilizando los parámetros del modelo de TRI como entrada para el análisis con el LSDM, mientras que los obtenidos en el Capítulo 5 reflejan errores de especificación en \mathbf{Q} , por lo tanto los valores MAD encontrados en el presente estudio, con datos reales, pueden deberse tanto a errores de especificación de \mathbf{Q} , como imprecisión del modelo de TRI.

Si se analizan los valores WMAD se puede observar que la recuperación es muy buena pues, como se esperaba, la ponderación minimiza el efecto de los errores en la estimación de las CPAs. Los valores RMSD son algo más altos que los del estadístico MAD pero en general son bastante similares. Para poder realizar una valoración precisa de la recuperación de las CCIs con estos dos nuevos estadísticos es necesario estudiar su distribución y encontrar puntos de corte adecuados al igual que se ha hecho con los valores MAD.

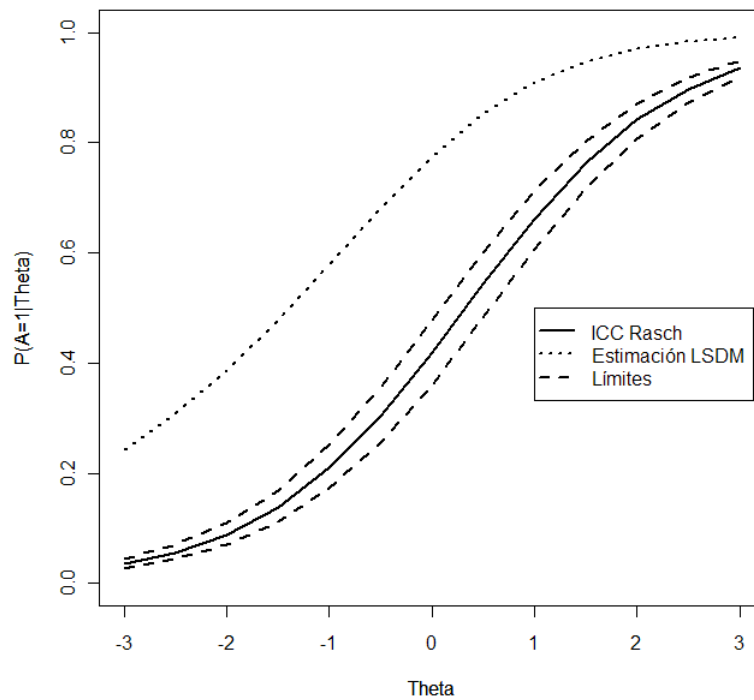
Tabla 6.5: Estadísticos de recuperación de las CCIs

Item	Q						Estadísticos de recuperación		
	A ₁	A ₃	A ₄	A ₅	A ₆	A ₆	WMAD	RMSD	MAD
1	1	0	0	0	0	0	.002	.098	.086
2	0	0	0	0	1	0	.007	.267	.246*
3	0	0	0	1	1	0	.006	.230	.209*
4	0	0	1	0	0	1	.005	.194	.175*
5	0	1	0	0	0	0	.000	.036	.023
6	0	0	0	0	0	1	.003	.122	.111*
7	1	0	0	0	0	0	.002	.090	.079
8	0	0	1	0	0	0	.000	.022	.020
9	0	0	0	0	0	1	.000	.015	.012
10	0	0	1	0	0	0	.001	.075	.063
11	0	0	1	0	0	0	.003	.137	.124*
12	0	0	0	1	0	0	.003	.111	.102*
13	0	0	0	0	1	0	.006	.257	.236*
14	1	0	0	0	0	1	.002	.099	.086*
15	0	0	1	0	0	0	.000	.023	.021
16	0	0	1	0	1	0	.000	.033	.030
17	0	0	0	0	1	0	.005	.221	.201*
18	0	0	0	1	0	0	.006	.236	.215*
19	0	0	0	1	1	0	.006	.233	.211*
20	0	0	0	0	0	1	.000	.024	.022
21	0	1	1	0	0	0	.002	.076	.070
22	0	0	0	0	0	1	.000	.013	.011
23	0	0	1	0	0	0	.003	.137	.121
25	0	0	1	0	0	0	.003	.129	.118*
26	0	0	0	0	0	1	.002	.080	.073
27	0	0	1	0	0	0	.000	.021	.018
28	0	0	1	0	0	0	.003	.114	.104*
29	0	0	1	0	1	0	.003	.113	.102*
30	0	0	1	0	0	0	.003	.111	.102*
31	0	1	1	0	0	0	.001	.044	.041
32	0	0	0	1	0	0	.002	.095	.086
33	0	1	0	0	0	0	.001	.083	.068
34	0	0	0	1	0	0	.003	.132	.121*
35	0	1	0	0	1	0	.003	.137	.121*
36	0	1	0	0	0	0	.000	.040	.025
37	0	0	1	0	0	0	.001	.062	.057
38	0	0	0	1	0	0	.003	.129	.119*
39	1	0	0	0	0	0	.001	.057	.049

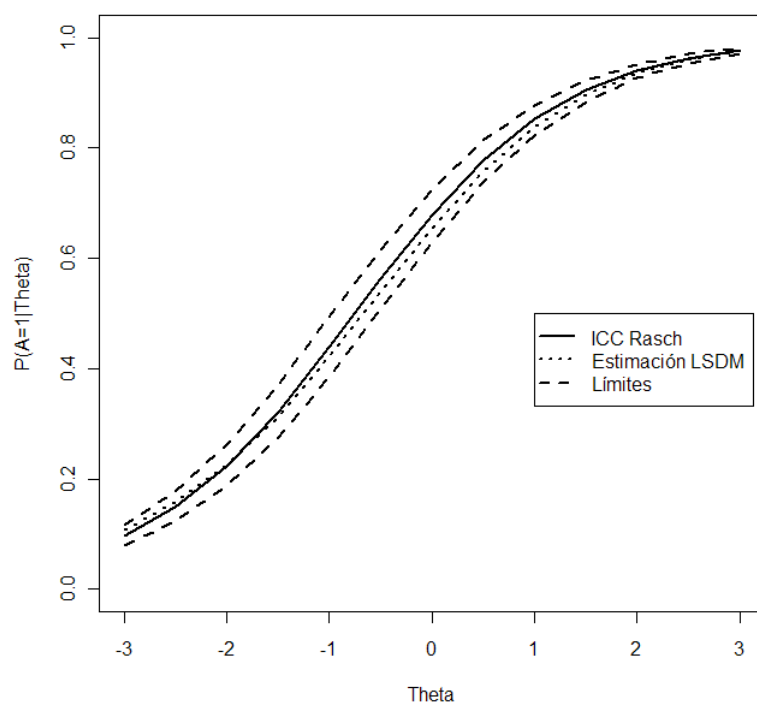
En las Gráficas 6.4 a 6.6 se representan las CCIs y límites del ítem 2 ($MAD = 0.184$), del ítem 8 ($MAD = 0.021$) y del ítem 16 ($MAD = 0.030$). En la Gráfica 6.4 se puede observar que la CCI estimada por el LSDM para el ítem 2 sobrepasa los límites en todos los niveles de habilidad, indicando que se trata de un ítem con recuperación muy pobre. La

Gráfica 6.5 representa la situación contraria pues se trata de un ítem muy bien recuperado cuya CCI se encuentra dentro de los límites establecidos. La Gráfica 6.6 ilustra la situación en la que la CCI estimada por el LSDM excede los límites en los niveles bajos de habilidad y se encuentra dentro en los niveles altos ($\theta > 0$). En este ejemplo se observa cómo la metodología de límites complementa los resultados de los estadísticos de recuperación (MAD, WMAD y RMSD) pues estos últimos, al constituir valores promedio en los niveles de habilidad, no permiten conocer si la recuperación es homogénea a lo largo de los diversos niveles de θ .

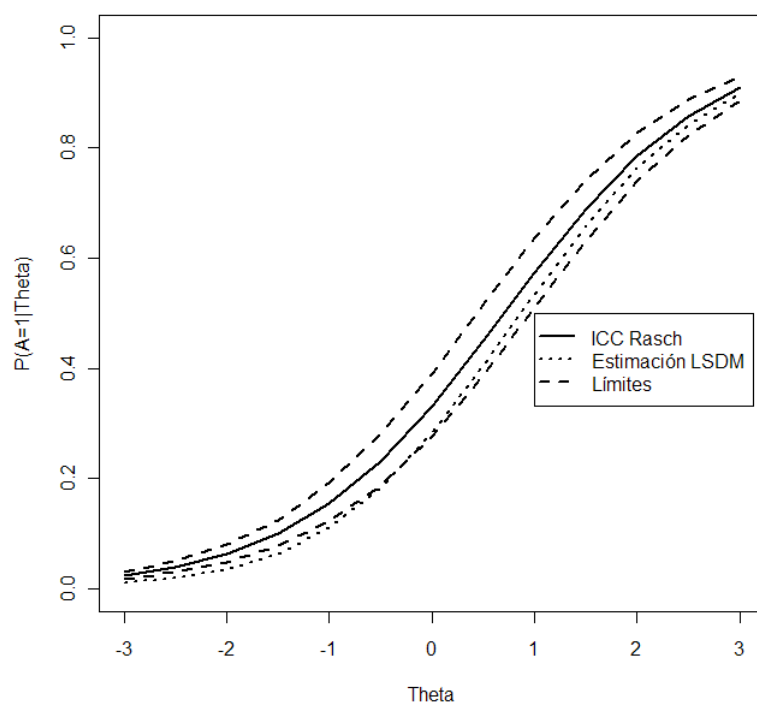
Gráfica 6.4: CCI original y recuperada con límites para el ítem 2



Gráfica 6.5: CCI original y recuperada con límites para el ítem 8



Gráfica 6.6: CCI original y recuperada con límites para el ítem 16



6.6.4. Validación cruzada de los resultados con los modelos LLTM y DINA

6.6.4.1. Modelo LLTM

La correlación entre los parámetros de los ítems estimados con el LLTM y el modelo de Rasch ha sido 0.84 y el estadístico LR indica desajuste: $\chi^2 = 49.6$ ($p < 0.01$) con 29 grados de libertad. A pesar de lo anterior se proseguirá con el análisis porque el objetivo es la validación cruzada de los resultados de acuerdo al orden de las dificultades de los atributos y no la predicción de las dificultades de los ítems. Adicionalmente, varios autores, entre ellos el mismo Fischer (1995) han notado que el estadístico LR es demasiado conservador indicando desajuste del modelo aún cuando la correlación entre los parámetros de dificultad sea alta. Los valores de los parámetros básicos, estimados con el LLTM, ordenados de fácil a difícil han sido: $\alpha_1 = -1.59$, $\alpha_2 = -0.05$, $\alpha_5 = 0.25$, $\alpha_6 = 1.03$, $\alpha_4 = 1.12$ y $\alpha_3 = 1.49$, coincidiendo perfectamente con el orden de las CPAs estimadas con el LSDM, adicionalmente todos han sido significativamente diferentes de 0 ($p < 0.01$) indicando que todos los atributos contribuyen a la explicación de la dificultad de los ítems.

6.6.4.2. Modelo DINA

El modelo DINA estima las probabilidades posteriores de que cada examinado pertenezca a cada una de las clases latentes conformadas por los 2^k patrones de dominio de las habilidades latentes, siendo k el número de atributos. En el caso del TA se conforman $2^6 = 64$ patrones de habilidad o clases latentes en las que son clasificados los estudiantes. La probabilidad de dominio de un atributo para un sujeto en DINA está dada por la suma de las probabilidades posteriores de pertenencia a cada clase en la que se supone dominio el atributo. La denominada prevalencia del atributo es la media de dichas probabilidades para cada atributo, y es una medida de la dificultad relativa del atributo. La prevalencia se ha utilizado como criterio para comparar el orden de la dificultad relativa de los atributos estimados en el DINA con el orden encontrado en los modelos LSDM y LLTM. El orden de las prevalencias estimadas por el modelo DINA ha sido: $A_1 = 0.78$, $A_2 = 0.62$, $A_3 = 0.35$, $A_4 = A_5 = 0.32$ y $A_6 = 0.24$. El resultado es algo diferente al evidenciado con los modelos LLTM y LSDM pues el atributo A_3 aparece situado en tercer lugar y no como el atributo

de mayor dificultad relativa. Adicionalmente el atributo de simplificación A_6 se ordena como el de mayor dificultad en lugar del atributo A_4 . Los demás atributos se ordenan de forma similar a la encontrada con los modelos LLTM y LSDM.

En la Tabla 6.6 se presenta un resumen de la comparación del ordenamiento de los atributos según su dificultad relativa entre los modelos DINA, LLTM, LSDM y con la proporción de dominio observada empíricamente, mediante la proporción media de acierto de los ítems individuales. El ordenamiento producido por los modelos LLTM y LSDM coincide perfectamente con lo esperado teóricamente y con la proporción de dominio empírica, siendo una importante evidencia de validez y consistencia de los resultados encontrados. Sin embargo, no ocurre lo mismo con el modelo DINA pues los atributos A_3 , A_4 , A_5 y A_6 aparecen en un orden diferente al esperado.

Tabla 6.6: Comparación del orden de la dificultad relativa de los atributos

DINA	$A_1 - A_2 - A_3 - A_4 - A_5 - A_6$
LLTM	$A_1 - A_2 - A_5 - A_6 - A_4 - A_3$
LSDM	$A_1 - A_2 - A_5 - A_6 - A_4 - A_3$
Observada	$A_1 - A_2 - A_5 - A_6 - A_4 - A_3$

6.6.5. Comparación de los estados de conocimiento obtenidos con LSDM y DINA con los patrones observados

La Tabla 6.7 contiene los estadísticos RMSD y MAD para analizar la magnitud de las diferencias entre las matrices de probabilidad de dominio de los atributos o *estados de conocimiento* obtenidas con los modelos de diagnóstico cognitivo (DINA y LSDM) y la proporción de acierto observada en los ítems individuales. Los estadísticos han sido calculados para cada atributo teniendo en cuenta los 382 estudiantes examinados. En dicha tabla se puede observar que el modelo DINA se aproxima con mayor precisión a los estados de conocimiento observados excepto para el caso de los atributos A_3 y A_5 para los cuales el LSDM presenta una mejor aproximación.

Tabla 6.7: Estadísticos MAD y RMSD para los modelos DINA y LSDM

	MAD		RMSD	
	LSDM	DINA	LSDM	DINA
A_1	0.229	0.098	0.260	0.200
A_2	0.279	0.132	0.333	0.215
A_3	0.138	0.265	0.172	0.376
A_4	0.239	0.159	0.276	0.261
A_5	0.247	0.257	0.298	0.323
A_6	0.297	0.111	0.344	0.189

6.7. Conclusiones

En el presente estudio se ilustra la aplicación del método LSDM (Dimitrov, 2007) en un proceso de validación y análisis de un conjunto de seis atributos cognitivos propuestos para la solución de tareas de adición y sustracción con números enteros. El comportamiento lógico y monótono de las CPAs proporciona una importante evidencia sobre la validez global de la estructura cognitiva planteada para el test de aritmética, adicionalmente, el ordenamiento de los atributos de acuerdo a su dificultad coincide con las hipótesis teóricas y con la proporción de acierto de los ítems individuales diseñados para medir el desempeño en los componentes. Sin embargo, es importante tener en cuenta que ha sido necesario realizar una reespecificación en la matriz \mathbf{Q} para eliminar una relación de dependencia secuencial entre los atributos A_1 y A_4 . Dicha reespecificación ha mejorado la solución obtenida con el método, como se ha podido observar en los valores residuales LSD y en el comportamiento lógico de las CPAs. La congruencia de los resultados, tras efectuar la reespecificación, demuestra la utilidad del método LSDM para la validación de atributos cognitivos, siempre y cuando los cambios realizados en la matriz \mathbf{Q} tengan un sentido teórico.

En cuanto a la recuperación de las CCIs se ha podido evidenciar que 18 de los 39 ítems no presentan una recuperación adecuada. Una explicación posible a estos resultados es que el TA sólo contiene una o dos tareas de cada tipo y puede que sea necesario tener más ítems que aporten mayor información sobre la descomposición en atributos y por tanto una mayor precisión en la recuperación de las CCIs. El presente trabajo ha permitido utilizar los puntos de corte encontrados en los estudios de simulación e ilustrar en una aplicación real los estadísticos de recuperación de las CCIs que se proponen en el presente estudio:

el MAD ponderado (WMAD), la raíz media de la diferencia al cuadrado (RMSD) y los límites para las CCIs utilizando el error estándar de los parámetros de dificultad obtenidos con la TRI. Las tres nuevas aproximaciones han demostrado ser útiles para la detección de errores de especificación en los vínculos entre ítems y atributos, proporcionando nuevas herramientas para la utilización práctica del LSDM.

En cuanto a la validación cruzada de resultados se ha encontrado un correcto emparejamiento en el ordenamiento de los atributos según los parámetros básicos del LLTM, las prevalencias de dominio de los atributos obtenidas con el modelo DINA, la proporción de acierto de los ítems individuales, la proporción media de dominio estimada por el LSDM y lo esperado por la teoría. Lo anterior constituye una importante evidencia de la consistencia de los resultados.

Con respecto a la capacidad diagnóstica del LSDM se ha encontrado que dicho método presenta una mayor congruencia con los estados de conocimiento observados en el atributo A_3 y A_5 mientras que el modelo DINA se acerca más a la clasificación observada en los restantes atributos. Sin embargo los valores RMSD y MAD no son excesivamente grandes indicando que el método LSDM puede ser utilizado para propósitos de diagnóstico cognitivo siempre y cuando se cuente con una estimación adecuada de las CPAs y con unas propiedades lógicas en cuanto al orden de la dificultad de los atributos.

Los resultados del presente trabajo ilustran la importancia de tener en cuenta componentes cognitivos en las primeras fases del proceso evaluativo (diseño del test) que permitan al investigador otorgar un significado más preciso a las puntuaciones obtenidas y ofrecer retroalimentación más específica a los examinados y, en este caso, a los profesores del área de interés (aritmética). La aproximación ilustrada en este estudio puede ser aplicada también en otras áreas del conocimiento y conducir a una mejora en los procesos de enseñanza-aprendizaje en dichas áreas.

Capítulo 7

Discusión General

En la presente tesis se han estudiado las propiedades y aplicaciones del método de las distancias mínimo cuadráticas (LSDM). Los modelos de diagnóstico cognitivo han cobrado especial relevancia en los últimos años porque han permitido comprender a fondo los procesos de medición y propenden por una mejoría de las dificultades diagnosticadas en los examinados y no sólo por alcanzar un estatus de aprendizaje. El objetivo de este enfoque es interpretativo, diagnóstico y prescriptivo llegando mucho más allá de la simple retroalimentación de la puntuación.

El enfoque de diagnóstico cognitivo comprende, no sólo un conjunto de modelos psicométricos alternativos, sino una concepción global de todo el proceso de medición, incluyendo como piedra angular la validez, mediante un entendimiento completo del constructo a evaluar desde las fases tempranas de la medición. Es por dicho motivo que la realización de un proceso de medida bajo esta perspectiva debe partir siempre de la especificación cognitiva del constructo a medir, dicha especificación requiere la participación activa de otras áreas del conocimiento externas a la Psicometría como la Psicología Cognitiva, la Educación, la Neuropsicología..., convirtiendo los procesos evaluativos en un esfuerzo interdisciplinario. El instrumento de medición debe ser el reflejo de este profundo estudio del constructo y sólo cuando se cumple dicha característica tiene sentido la aplicación de los modelos diagnósticos pues los resultados de la medición serán útiles y enriquecedores.

La mayoría de los modelos componenciales se centran en el diagnóstico de la presencia o ausencia de múltiples variables latentes de tipo discreto llamadas atributos (e.g DINA; Junquer y Sijtsma, 2001; RSM, Tatsuoaka, 1995; RUM, DiBello, Stout y Roussos, 1995).

Los atributos difieren en cuanto al grado de profundidad o *granularidad* y los múltiples estudios que se han basado en este enfoque analizan componentes tan variados como operaciones, reglas, procesos cognitivos y competencias, haciendo falta una discusión mucho mas profunda sobre qué es un atributo, cuál debe ser su nivel de profundidad, y cuáles son los modelos mas apropiados para reflejar sus relaciones, pues no es lo mismo analizar, como en la presente tesis, un conjunto de reglas u operaciones aritméticas que complejos procesos meta-cognitivos como la representación de número o la cardinalidad.

Dentro de los métodos propuestos para el análisis de componentes cognitivos se encuentra el LSDM (Dimitrov 2007). Dicho método ha sido originalmente propuesto para la validación y análisis de los atributos expresados en la matriz \mathbf{Q} , el LSDM asume una relación conjuntiva entre los atributos y estima las probabilidades de dominio de los mismos utilizando el álgebra lineal. A diferencia de otros modelos de diagnóstico cognitivo, en el LSDM ha sido diseñado para ser usado con medidas unidimensionales cuando se tienen parámetros de los ítems estimados con la TRI. Es un método parsimonioso, que no requiere complejos procesos de estimación de parámetros y es fácilmente interpretable porque las CPAs se expresan en la misma escala que las CCIs.

En el Capítulo 3 de la presente tesis se ha estudiado la recuperación de las CCIs en nueve condiciones experimentales utilizando procedimientos de simulación y se han generalizado puntos de corte para la valoración de la calidad de la recuperación de las CCIs. Para cumplir dicho objetivo se ha determinado la distribución empírica de los valores MAD encontrando que el aspecto fundamental en la calidad de la recuperación es el número de atributos requeridos por los ítems pues la recuperación es mucho mejor en ítems que requieren por lo menos dos atributos. En este estudio también se han encontrado puntos de corte para valorar la recuperación de las CCIs en contextos aplicados. El aporte principal de dicho trabajo es que permite al investigador tener un criterio objetivo para reespecificar los vínculos de ítems particulares en la matriz \mathbf{Q} , siendo una herramienta importante en el proceso de validación de la estructura cognitiva. Al desconocer la distribución del estadístico MAD y su comportamiento frente a diversas condiciones este criterio tenía un carácter subjetivo; con este estudio se ha ganado en la profundidad del conocimiento de dicha herramienta. El procedimiento propuesto en este trabajo también es novedoso porque no se parte de la generación de datos con respuestas de los examinados al test sino

que estas últimas se generan a partir de los estados de conocimiento (Tatsuoka, 1995).

En los Capítulos 4 y 5 se estudian los efectos de la introducción de errores de especificación en la matriz \mathbf{Q} sobre los criterios propuestos por Dimitrov (2007) para la validación de atributos cognitivos con el LSDM. En concreto se quiso determinar si dichos criterios son sensibles a la presencia de error y en qué grado. Se ha buscado que los errores introducidos en la estructura cognitiva de estos estudios tengan un sentido teórico, en otras palabras, que sean equivocaciones que pueden surgir en cualquier aplicación real de este tipo de mediciones. Por dicho motivo se ha procurado abarcar posibles fuentes de error reportadas en estudios empíricos y en las investigaciones previas en las que también se han estudiado los efectos de errores de especificación en la matriz \mathbf{Q} (Rupp y Templin, 2008; Baker, 1993; de la Torre, 2008). Además, se ha evitado introducir errores al azar.

Adicionalmente, para la realización de estos estudios no se han usado los parámetros de la TRI sino que se ha analizado directamente la matriz de probabilidad de acierto de los ítems (P_{ij}) obtenida mediante la multiplicación de las probabilidades de dominio de los atributos generadas con el modelo de Rasch. Siguiendo este procedimiento se ha buscado valorar únicamente el efecto de la introducción de errores de especificación en la estructura cognitiva sin otras fuentes de error como la estimación de los parámetros de la TRI o el procedimiento de simulación de respuestas a los ítems.

Se ejecutó primero un estudio descriptivo en el que se examinaban los efectos de cuatro tipos de error que afectaban tanto a la matriz \mathbf{Q} como al comportamiento de los atributos. Los resultados de este trabajo han mostrado que los diversos tipos de error de especificación producen efectos importantes en los criterios del método LSDM, tanto a nivel de la estimación de las probabilidades de dominio de los atributos como en la recuperación de las CCIs de los ítems. En primer lugar, los errores en los vínculos entre ítems y atributos producen un empeoramiento de la recuperación de las CCIs, en especial cuando se modifican los vínculos de los ítems con los atributos de mayor dificultad relativa. La fusión entre atributos difíciles, las incompatibilidades y la heterogeneidad de los atributos también producen un empobrecimiento de la recuperación de las CCIs. A pesar de los resultados anteriores resulta muy difícil conocer el ítem particular que se encuentra mal especificado pues algunos ítems que no han sido objeto de ningún cambio presentan también una pobre recuperación.

En cuanto a los valores LSD se ha encontrado que los errores de especificación producen un incremento en los residuales del método, especialmente en los niveles bajos de habilidad, confirmando lo encontrado en otras aplicaciones del LSDM.

Otro de los resultados importantes de este estudio es que la estimación de las probabilidades de dominio de los atributos se han visto afectadas por la presencia de errores de especificación en la estructura cognitiva. Es interesante observar que la fusión, incompatibilidades y errores en los vínculos de los atributos difíciles producen cambios en las curvas de probabilidad de los atributos fáciles. En particular ocurre una sobrestimación de las probabilidades de dominio de dichos atributos llegando incluso a producir probabilidades constantes de 1 para todos los niveles de habilidad. Este es un resultado común en los estudios aplicados del LSDM (Ma, Çetin y Green, 2009; Dimitrov, 2007), incluyendo los estudios de aplicación en un test de aritmética presentados en el Capítulo 6. Una explicación plausible a dicho resultado es que la solución del método para el vector \mathbf{X} asociado al atributo en cuestión es nula y al ser multiplicado por la matriz \mathbf{Q} no altera las restantes probabilidades, actuando como módulo en la multiplicación de \mathbf{QX} . Como las probabilidades de dominio de los atributos son el exponente del vector \mathbf{X} la curva de probabilidad es una constante de 1, sin embargo, este resultado no se puede interpretar como una estimación real del dominio del atributo sino que es un artefacto matemático que en realidad significa que el atributo en cuestión no aporta información a la habilidad para resolver los ítems (al ser nula su estimación no está participando en la descomposición de los ítems). Este resultado se hace aún más evidente cuando se adiciona un atributo originalmente no contemplado en la estructura cognitiva pues las probabilidades de dominio estimadas para el atributo adicionado son una constante de 1 en todos los niveles de habilidad y las probabilidades de los atributos restantes no se ven modificadas produciendo valores MAD de 0 y una recuperación exacta de las CCIs.

Frente a este estudio cabe discutir dos resultados más: a) no se ha evidenciado faltas de monotonidad en ninguna de las curvas a pesar de la introducción de los errores de especificación y b) la heterogeneidad de los atributos, es decir, el hecho de que las dificultades de los atributos sean diferentes dependiendo del ítem en el que se aplique el atributo afecta la estimación de las CPAs. El primer resultado conduciría a replantear la efectividad de uno de los criterios de validación propuestos para este método por su autor

y el segundo previene de la importancia de cumplir el supuesto de homogeneidad en los atributos, es decir, para realizar aplicaciones prácticas del LSDM se debe procurar que un atributo k tenga siempre la misma dificultad, independientemente del ítem en el que sea utilizado.

El segundo estudio que aborda los efectos de los errores de especificación de la estructura cognitiva se ha realizado con el objeto de generalizar algunos de los resultados observados en el estudio previo descriptivo y ampliar las condiciones de simulación. Para lograr dicho objetivo se analizaron las propiedades estadísticas, en particular se estudió la sensibilidad y especificidad de los valores MAD y LSD para detectar la presencia de 14 tipos de error en la matriz \mathbf{Q} .

A nivel metodológico fue necesario idear un procedimiento de aleatorización que permitiera obtener réplicas de la matriz P_{ij} sin tener que calibrar diversas matrices de datos con el modelo de TRI puesto que se quería tener una estimación de los efectos debidos únicamente a la introducción de errores de especificación y no a los procesos de simulación y calibración. Se decidió, entonces, utilizar una distribución binomial para asignar probabilidades ligeramente diferentes a la matriz P_{ij} original y obtener de dicha forma 200 matrices replicadas con las que se hicieron los análisis de sensibilidad y especificidad de los dos estadísticos de interés.

Los resultados del estudio muestran que el estadístico MAD es sensible a la detección de errores de especificación introducidos en atributos con mayor dificultad relativa, como se esperaba, dicho resultado concuerda con los del estudio descriptivo presentado en el Capítulo 4, sin embargo, el estadístico no es sensible a errores que afectan los atributos fáciles. Por otra parte el estadístico LSD no es sensible a la detección de ninguno de los tipos de error de especificación de la matriz \mathbf{Q} analizados en dicho estudio. Otro resultado interesante es que el comportamiento de los criterios del LSDM es constante a lo largo de las diversas condiciones de simulación.

En el Capítulo 6 se presentan las aplicaciones prácticas del método y se proponen varias extensiones de los criterios de evaluación de la validez de los atributos cognitivos. También se propone una metodología para usar el LSDM como herramienta para el diagnóstico cognitivo. En primer lugar se ha elaborado un instrumento para la medición de operaciones de adición y sustracción entre números enteros basado en una estructura cog-

nitiva profundamente estudiada y compuesta por seis atributos. Dicho instrumento ha sido diseñado para conocer tanto el desempeño de los examinados en cada una de las tareas como la ejecución de cada uno de los pasos realizados para resolverlas. El test también se ha basado en un instrumento previo, de operaciones con enteros (Real y cols, 1999).

Por medio de la validación y análisis de dicha estructura se ha buscado ilustrar la aplicación de los resultados obtenidos en los capítulos precedentes y valorar la utilidad de las nuevas técnicas propuestas para el análisis de la recuperación de las CCI's. El análisis inicial de los atributos pone en evidencia una relación de dependencia secuencial entre una pareja de atributos por lo que se plantea una reespecificación de la estructura cognitiva. Dicha reespecificación se ha basado no sólo en los resultados del LSDM sino que también tiene una base lógica dentro de la teoría cognitiva que sustenta la estructura propuesta. Los resultados del método, tras la reespecificación, aportan evidencia sobre la validez de la estructura y refuerzan la utilidad de los criterios del método para evaluar los atributos propuestos.

Para contrastar los resultados se ha realizado también una validación cruzada del ordenamiento de las dificultades de los atributos con otros dos modelos componenciales (LLTM y DINA) y con los resultados empíricos de la ejecución de los examinados en cada uno de los atributos individuales encontrando resultados coherentes entre las 4 aproximaciones y lo esperado según la teoría cognitiva en la que se sustenta el estudio.

En este trabajo se han propuesto tres formas alternativas de evaluar la calidad de la recuperación de las CCI's con el LSDM, dichas técnicas pueden complementar los valores MAD y proporcionar nueva evidencia al respecto. Especialmente útil ha resultado la metodología basada en los límites construidos alrededor de la CCI pues permiten valorar la calidad de la recuperación en los diferentes niveles de habilidad ya que con los valores MAD, WMAD y RMSD, al ser promedios de diferencias en los diversos niveles de habilidad, se pierde dicha información que puede ser interesante y relevante tener en cuenta ya que un ítem particular puede estar muy bien recuperado en ciertos niveles y muy mal en otros (especialmente en niveles bajos de habilidad donde el método LSDM es más impreciso).

Con respecto a la capacidad del método propuesto para utilizar el LSDM como herramienta para el diagnóstico cognitivo se ha evidenciado que resulta útil siempre y cuando la

estimación de las CPAs sea lógica y corresponda con lo esperado teóricamente. La metodología propuesta depende de la estimación de los parámetros de habilidad según la TRI, entonces, otro requisito para su aplicación es la precisión en dicha estimación y el correcto ajuste del modelo de TRI.

Los resultados de este estudio ilustran la importancia de tener en cuenta componentes cognitivos desde las primeras fases del proceso evaluativo y también en el diseño de los instrumentos de medida. Muchas aplicaciones de los modelos componenciales se han realizado con datos empíricos provenientes de test que fueron originalmente desarrollados desde una perspectiva clásica (sumativa) más que con propósitos de diagnóstico cognitivo, consecuentemente, el número de aplicaciones prácticas exitosas de este tipo de modelos es bastante bajo ocurriendo con frecuencia problemas de convergencia, limitaciones en el número de atributos, bajas fiabilidades en las subescalas, valores altos en los parámetros estimados, entre otros problemas (e.g Sinharay, Almond y Yan, 2004; Dibello, Stout y Roussos, 1995; Romero, Ponsoda y Ximénez, 2008).

Independientemente de los aportes y limitaciones de los presentes estudios es inmenso el potencial futuro de este tipo de aproximaciones y también muy grande la cantidad de trabajo que hay por delante. Sin ánimo de exhaustividad, algunos de los tópicos que se pueden seguir investigando son la realización de extensiones del método a ítems no binarios (politómicos), el análisis del comportamiento del LSDM frente a otras formas de representar la estructura cognitiva en la matriz \mathbf{Q} , no necesariamente binarias, por ejemplo matrices \mathbf{Q} de carácter politómico, o de frecuencias de uso de los atributos, la aplicación del método a otras áreas donde exista un estudio más profundo de los procesos cognitivos desde la Psicología Cognitiva pues, si bien el área de la aritmética mental está muy estudiada, no ocurre lo mismo con las operaciones entre números enteros, el estudio de las nuevas herramientas propuestas para valorar la calidad de la recuperación de las CCIs (WMAD, RMSD y límites), estudiar el efecto conjunto en el indicador MAD de introducir errores en \mathbf{Q} y la aplicación de la TRI, analizar los efectos de la multidimensionalidad del test en el método LSDM y extender el método para incluir otro tipo de relaciones entre los atributos, por ejemplo, disyuntivas.

A nivel de los modelos componenciales se han encontrado diversidad de modelos y cada año se proponen más, sin embargo, las aplicaciones empíricas muy son escasas, por tanto,

en el futuro se pueden empezar a utilizar más este tipo de modelos que enriquecen ampliamente los resultados de la medición. Tendría por tanto interés explorar las aplicaciones en el campo educativo y no educativo de estos modelos.

Para finalizar, en la presente tesis se ha presentado el estudio de las propiedades y aplicaciones del método LSDM, independientemente de los aportes y limitaciones que se han detectado en el método, los estudios que configuran el presente trabajo proporcionan información sobre las condiciones bajo las cuales el potencial teórico del LSDM puede ser explotado en la práctica de la medición.

Capítulo 8

Anexos

8.1. Anexo 1

Matrices **Q** y percentil 95 del análisis de recuperación de las CCIs (Capítulo 3).

Tabla 8.1: Matriz **Q** y percentil 95 en la condición 6A15I

	A_1	A_2	A_3	A_4	A_5	A_6	P_{95}
i1	0	0	1	0	0	0	.0381
i2	1	0	0	0	0	1	.0163
i3	0	1	0	1	0	0	.0545
i4	0	0	1	1	0	0	.0176
i5	0	1	1	0	1	0	.0140
i6	1	0	1	0	0	1	.0484
i7	0	1	0	0	1	1	.0291
i8	0	0	1	1	0	1	.0106
i9	1	0	1	0	1	0	.0212
i10	0	1	1	1	0	1	.0270
i11	1	1	0	0	1	1	.0177
i12	1	1	0	1	0	1	.0279
i13	1	1	0	1	1	0	.0098
i14	1	0	1	1	1	1	.0093
I15	1	1	1	1	1	1	.0140

Tabla 8.2: Matriz \mathbf{Q} y percentil 95 en la condición 8A15I

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	P_{95}
i1	0	0	0	1	0	0	0	0	.0603
i2	0	1	0	0	0	0	1	0	.0549
i3	1	0	1	0	1	0	0	0	.0179
i4	0	0	1	1	0	1	0	0	.0356
i5	0	1	0	1	0	0	1	0	.0495
i6	1	0	0	1	1	0	0	1	.0140
i7	0	0	1	1	1	0	1	0	.0207
i8	0	0	1	0	0	1	1	1	.0384
i9	0	1	1	0	0	1	1	0	.0391
i10	1	0	0	1	1	0	1	1	.0205
i11	1	1	0	1	0	1	0	1	.0097
i12	0	1	1	0	1	0	1	1	.0379
i13	1	1	1	0	1	1	0	1	.0132
i14	1	1	0	1	1	1	1	1	.0085
i15	1	1	1	1	1	1	1	1	.0140

Tabla 8.3: Matriz \mathbf{Q} y percentil 95 en la condición 4A24I

	A_1	A_2	A_3	A_4	P_{95}
i1	1	0	0	0	.0703
i2	0	1	0	0	.0692
i3	0	0	1	0	.0614
i4	0	0	0	1	.0730
i5	0	1	0	0	.0692
i6	0	0	1	0	.0614
i7	1	1	0	0	.0613
i8	1	0	1	0	.0373
i9	1	0	0	1	.0375
i10	0	1	1	0	.0212
i11	0	1	0	1	.0201
i12	0	0	1	1	.0290
i13	1	0	1	0	.0281
i14	1	0	0	1	.0373
i15	0	1	1	0	.0375
i16	1	1	1	0	.0201
i17	1	1	0	1	.0181
i18	1	0	1	1	.0208
i19	0	1	1	1	.0121
i20	1	1	1	0	.0141
i21	1	1	0	1	.0181
i22	1	1	0	1	.0208
i23	1	1	1	1	.0230
i24	1	1	1	1	.0231

Tabla 8.4: Matriz \mathbf{Q} y percentil 95 en la condición 6A24I

	A_1	A_2	A_3	A_4	A_5	A_6	P_{95}
i1	0	1	0	0	0	0	.1179
i2	0	0	0	0	1	0	.0848
i3	1	0	0	1	0	0	.0793
i4	0	1	0	0	1	0	.0590
i5	0	0	1	1	0	0	.0545
i6	0	0	0	1	0	1	.0334
i7	0	0	0	0	1	1	.0699
i8	1	1	0	0	0	1	.0445
i9	1	0	1	0	0	1	.0355
i10	1	0	0	0	1	1	.0429
i11	0	1	1	1	0	0	.0523
i12	0	1	0	1	1	0	.0368
i13	0	1	0	0	1	1	.0138
i14	0	0	1	1	1	0	.0163
i15	0	0	1	0	1	1	.0167
i16	0	0	0	1	1	1	.0271
i17	1	1	1	1	0	0	.0540
i18	1	1	1	0	0	1	.0118
i19	1	0	0	1	1	1	.0121
i20	0	1	1	1	1	0	.0108
i21	0	1	0	1	1	1	.0267
i22	1	1	1	1	0	1	.0268
i23	1	1	1	0	1	1	.0223
i24	1	1	1	1	1	1	.0306

Tabla 8.6: Matriz \mathbf{Q} y percentil 95 en la condición 4A32I

	A_1	A_2	A_3	A_4	P_{95}
i1	1	0	0	0	.0685
i2	0	1	0	0	.0744
i3	0	0	1	0	.0614
i4	0	0	0	1	.0716
i5	1	0	0	0	.0685
i6	0	1	0	0	.0744
i7	0	0	1	0	.0614
i8	0	0	0	1	.0716
i9	1	1	0	0	.0631
i10	1	0	1	0	.0366
i11	1	0	0	1	.0301
i12	0	1	1	0	.0291
i13	0	1	0	1	.0279
i14	0	0	1	1	.0302
i15	1	1	0	0	.0631
i16	1	0	1	0	.0366
i17	1	0	0	1	.0301
i18	0	1	1	0	.0291
i19	0	1	0	1	.0279
i20	0	0	1	1	.0302
i21	1	0	1	0	.0366
i22	1	1	1	0	.0210
i23	1	1	0	1	.0308
i24	1	0	1	1	.0175
i25	0	1	1	1	.0156
i26	1	1	1	0	.0210
i27	1	1	0	1	.0308
i28	1	0	1	1	.0175
i29	0	1	1	1	.0156
i30	1	1	1	1	.0294
i31	1	1	1	1	.0294
i32	1	0	0	1	.0301

Tabla 8.7: Matriz \mathbf{Q} y percentil 95 en la condición 6A32I

	A_1	A_2	A_3	A_4	A_5	A_6	P_{95}
i1	1	0	0	0	0	0	.1131
i2	0	0	0	1	0	0	.0977
i3	0	0	0	0	0	1	.1223
i4	1	0	1	0	0	0	.0537
i5	1	0	0	0	1	0	.0646
i6	0	1	1	0	0	0	.0823
i7	0	1	0	0	0	1	.0591
i8	0	0	1	0	0	1	.0431
i9	0	0	0	1	1	0	.0638
i10	0	0	0	0	1	1	.1031
i11	1	1	1	0	0	0	.0897
i12	1	1	0	1	0	0	.0372
i13	1	0	1	1	0	0	.0310
i14	1	0	1	0	1	0	.0490
i15	1	0	0	1	1	0	.0315
i16	1	0	0	1	0	1	.0225
i17	0	1	1	1	0	0	.0454
i18	0	1	1	0	1	0	.0334
i19	0	1	0	1	1	0	.0256
i20	0	1	0	1	0	1	.0390
i21	0	0	1	1	0	1	.0221
i22	1	1	1	0	1	0	.0248
i23	1	1	1	0	0	1	.0229
i24	1	1	0	1	1	0	.0126
i25	1	0	1	0	1	1	.0256
i26	0	1	1	1	1	0	.0104
i27	0	1	0	1	1	1	.0126
i28	0	0	1	1	1	1	.0167
i29	1	1	1	1	1	0	.0148
i30	1	1	0	1	1	1	.0218
i31	0	1	1	1	1	1	.0207
i32	1	1	1	1	1	1	.0277

8.2. Anexo 2

Matrices \mathbf{Q} *verdaderas* sobre las que se han introducido los errores de especificación (Capítulo 5).

Tabla 8.9: Matriz \mathbf{Q} en la condición 4A15I

	A_1	A_2	A_3	A_4
i1	1	0	0	0
i2	0	1	0	0
i3	0	0	1	0
i4	0	0	0	1
i5	1	1	0	0
i6	1	0	1	0
i7	1	0	0	1
i8	0	1	1	0
i9	0	1	0	1
i10	0	0	1	1
i11	0	1	1	1
i12	1	0	1	1
i13	1	1	0	1
i14	1	1	1	0
I15	1	1	1	1

Tabla 8.10: Matriz \mathbf{Q} en la condición 4A32I

	A_1	A_2	A_3	A_4
i1	1	0	0	0
i2	0	1	0	0
i3	0	0	1	0
i4	0	0	0	1
i5	1	1	0	0
i6	1	0	1	0
i7	1	0	0	1
i8	0	1	1	0
i9	0	1	0	1
i10	0	0	1	1
i11	0	1	1	1
i12	1	0	1	1
i13	1	1	0	1
i14	1	1	1	0
i15	1	1	1	1
i16	1	0	0	0
i17	0	1	0	0
i18	0	0	1	0
i19	0	0	0	1
i20	1	1	0	0
i21	1	0	1	0
i22	1	0	0	1
i23	0	1	1	0
i24	0	1	0	1
i25	0	0	1	1
i26	0	1	1	1
i27	1	0	1	1
i28	1	1	0	1
i29	1	1	1	0
i30	1	1	1	1
i31	1	0	0	1
i32	1	1	1	0

Tabla 8.11: Matriz \mathbf{Q} en la condición 8A15I

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
i1	0	0	0	1	0	0	0	0
i2	0	1	0	0	0	0	1	0
i3	1	0	1	0	1	0	0	0
i4	0	0	1	1	0	1	0	0
i5	0	1	0	1	0	0	1	0
i6	1	0	0	1	1	0	0	1
i7	0	0	1	1	1	0	1	0
i8	0	0	1	0	0	1	1	1
i9	0	1	1	0	0	1	1	0
i10	1	0	0	1	1	0	1	1
i11	1	1	0	1	0	1	0	1
i12	0	1	1	0	1	0	1	1
i13	1	1	1	0	1	1	0	1
i14	1	1	0	1	1	1	1	1
i15	1	1	1	1	1	1	1	1

Tabla 8.12: Matriz \mathbf{Q} en la condición 8A32I

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
i1	1	1	1	1	1	1	1	1
i2	1	1	1	1	0	1	1	1
i3	1	1	0	1	1	1	1	1
i4	1	1	1	1	0	0	1	1
i5	0	1	1	1	1	1	0	1
i6	1	1	1	1	1	0	0	1
i7	1	0	0	1	1	1	1	0
i8	0	1	1	0	0	1	1	1
i9	1	0	1	1	0	0	1	1
i10	1	1	1	1	1	0	0	0
i11	1	1	1	1	0	0	0	1
i12	0	1	1	1	0	0	1	1
i13	0	1	0	1	1	0	1	0
i14	1	1	0	0	1	0	1	0
i15	1	0	1	0	0	1	1	0
i16	0	0	1	0	1	0	1	1
i17	0	1	0	1	1	1	0	0
i18	1	0	1	0	1	1	0	0
i19	0	0	1	0	1	1	0	1
i20	0	1	1	1	0	0	0	1
i21	0	1	0	0	1	0	1	1
i22	0	0	1	0	0	1	0	1
i23	1	1	0	1	0	0	0	0
i24	0	1	0	0	1	0	1	0
i25	0	0	1	1	1	0	0	0
i26	1	0	0	0	1	1	0	0
i27	0	0	0	0	1	1	0	1
i28	0	1	0	0	0	0	0	1
i29	0	1	0	0	1	0	0	0
i30	0	0	0	0	1	0	1	0
i31	1	0	0	0	0	0	0	0
i32	0	0	1	0	0	0	0	0

8.3. Anexo 3

Tabla 8.13: Test de Aritmética

Nombre:

Marca con una X la respuesta correcta como se indica en el ejemplo.

Prueba de Signos

Ejemplo: 2 - 5 = a) 7 b) 3 <input checked="" type="checkbox"/> c) -3 d) -7	3 - 7 = a) -3 + 7 b) -3 - 7 c) -(7 + 3) d) -(7 - 3)	0 + (-6) = a) 6 b) 0 c) -0 d) -6	-7 - (-9) = a) 16 b) 2 c) -2 d) -16	-9 - 7 = a) -(9 + 7) b) -(9 - 7) c) 9 + 7 d) 9 - 7
2 + 8 = a) 10 b) 6 c) -6 d) -10	8 + (-4) = a) -8 - 4 b) 8 - 4 c) -4 - 8 d) 4 - 8	2 - (-7) = a) -2 + 7 b) -2 - 7 c) 2 + 7 d) 2 - 7	-6 + 3 = a) -6 - 3 b) 6 - 3 c) -3 - 6 d) 3 - 6	2 - 0 = a) 2 b) 0 c) -1 d) -2
-9 - (-5) = a) -9 + 5 b) -9 - 5 c) 9 + 5 d) 9 - 5	-6 + 3 = a) 9 b) 3 c) -3 d) -9	-6 - 9 = a) 15 b) 3 c) -3 d) -15	-2 + (-8) = a) -2 + 8 b) -2 - 8 c) 2 + 8 d) 2 - 8	-6 - 0 = a) -(6 + 0) b) 6 + (-0) c) 6 + 0 d) 6 - 0
-2 + (-8) = a) 10 b) 6 c) -6 d) -10	-5 + 7 = a) -5 - 7 b) 5 - 7 c) -7 - 5 d) 7 - 5	-9 + (-7) = a) 16 b) 2 c) -2 d) -16	5 - 9 = a) -5 + 9 b) -5 - 9 c) -(9 + 5) d) -(9 - 5)	8 + (-4) = a) 12 b) 4 c) -4 d) -12
-9 - (-5) = a) 14 b) 4 c) -4 d) -14	-6 - 9 = a) -(6 + 9) b) -(6 - 9) c) 6 + 9 d) 6 - 9	0 + (-6) = a) -0 - 6 b) 0 - 6 c) -6 - 0 d) 6 - 0	-9 + 5 = a) -9 - 5 b) 9 - 5 c) -5 - 9 d) 5 - 9	8 - 4 = a) 12 b) 4 c) -4 d) -12
7 - 5 = a) 12 b) 2 c) -2 d) -12	-7 - (-9) = a) -7 + 9 b) -7 - 9 c) 7 + 9 d) 7 - 9	-9 + 9 = a) 18 b) 0 c) -1 d) -18	3 + (-7) = a) 10 b) 4 c) -4 d) -10	3 - 6 = a) -3 + 6 b) -3 - 6 c) -(6 + 3) d) -(6 - 3)
-9 + (-7) = a) -9 - 7 b) 9 + 7 c) -9 + 7 d) 9 - 7	2 - (-7) = a) 9 b) 5 c) -5 d) -9	3 + (-7) = a) -3 - 7 b) 3 - 7 c) -7 - 3 d) 7 - 3	-7 + 9 = a) -7 - 9 b) 7 - 9 c) -9 - 7 d) 9 - 7	-2 - 8 = a) -(2 + 8) b) -(2 - 8) c) 2 + 8 d) 2 - 8
9 + 7 a) 16 b) 2 c) -2 d) -16	-9 + 9 = a) -9 - 9 b) 9 - 9 c) 9 + 9 d) -9 + (-9)	5 - 7 = a) 12 b) 2 c) -2 d) -12	-5 + 7 = a) 12 b) 2 c) -2 d) -12	2 + 7 = a) 9 b) 5 c) -5 d) -9

Bibliografía

- AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, D.C.: Author.
- Agresti, A. (2002). *Categorical data analysis*. New Jersey: Wiley and sons.
- Almond, R. G., DiBello, L. V., Moulder, B., & Zapata-Rivera, J. (2007). Modeling diagnostic assessments with bayesian networks. *Journal of Educational Measurement*, 44(4), 341-359.
- Anderson, J. R. (1990). Analysis of student performance with the LISP tutor. En N. Frederiksen, R. Glaser, A. Lesgold & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition*. (pp. 27-50). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.
- Anderson, J. R., Greeno, J. G., Reder, L. M., & Simon, H. A. (2000). Perspectives on learning, thinking, and activity. *Educational Researcher*, 29(4), 11-13.
- Arabie, P., & Carroll, J. D. (1980). MAPCLUS: A mathematical programming approach to fitting the ADCLUS model. *Psychometrika*, 45(2), 211-235.
- Baker, F. B. (1993). Sensitivity of the linear logistic test model to misspecification of the weight matrix. *Applied Psychological Measurement*, 17(3), 201-210.
- Baker, F. B., & Kim, S. H. (2004). *Item response theory*. New York: Marcel Dekker.
- Bechger, T. M., Verstralen, H. H. F. M., & Verhelst, N. D. (2002). Equivalent linear logistic test models. *Psychometrika*, 67(1), 123-136.
- Bejar, I. I. (1993). A generative approach to psychological and educational measurement. En N. Frederiksen, R. J. Mislevy & I. I. Bejar (Eds.), *Test theory for a new generation of tests*. (pp. 323-357). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

- Bejar, I. I. (2002). Generative testing: From conception to implementation. En S. H. Irvine, & P. C. Kyllonen (Eds.), *Generating items for cognitive tests: Theory and practice., nov 1998, princeton, NJ, US* (pp. 199-217). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.
- Beland, A., & Mislevy, R. J. (1992). *Probability-based inference in a domain of proportional reasoning tasks*. ETS Technical Report. No. ETS-RR-92-15-ONR.
- Bermejo, V. (1990). *El niño y la aritmética*. Barcelona: Paidós Educador.
- Birenbaum, M., Tatsuoaka, C., & Xin, T. (2005). Large-scale diagnostic assessment: Comparison of eighth graders' mathematics performance in the united states, singapore and israel. *Assessment in Education: Principles, Policy & Practice*, 12(2), 167-181.
- Borsboom, D., & Mellenbergh, G. J. (2007). Test validity in cognitive assessment. En J. P. Leighton, & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications*. (pp. 85-115). New York, NY, US: Cambridge University Press.
- Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. Oxford, England: John Wiley and Sons.
- Butter, R., De Boeck, P., & Verhelst, N. (1998). An item response model with internal restriction on item difficulty. *Psychometrika*, 63(1), 47-63.
- Campbell, J. I. D. (. (2005). *Handbook of mathematical cognition*. New York, NY, US: Psychology Press.
- Carpenter, T. P. (1985). How children solve simple word problems. *Education and Urban Society*, 17(4), 417-425.
- Carpenter, T. P., & Moser, J. M. (1984). The acquisition of addition and subtraction concepts in grades one through three. *Journal for Research in Mathematics Education*, 15(3), 179-202.
- Carroll, John S. & Payne, J. W. (1976). *Cognition and social behavior*. Oxford, England: Lawrence Erlbaum.
- Corbett, A. T., Anderson, J. R., & O'Brien, A. T. (1995). Student modeling in the ACT programming tutor. En P. D. Nichols, S. F. Chipman & R. L. Brennan (Eds.), *Cognitively diagnostic assessment*. (pp. 19-41). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

- Corter, J. E. (1995). Using clustering methods to explore the structure of diagnostic tests. En P. D. Nichols, S. F. Chipman & R. L. Brennan (Eds.), *Cognitively diagnostic assessment*. (pp. 305-326). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.
- Corter, J. E., & Tversky, A. (1986). Extended similarity trees. *Psychometrika*, 51(3), 429-451.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281-302.
- de la Torre, J. (2008). An Empirically Based Method of Q-Matrix Validation for the DINA Model: Development and Applications, *Journal of Educational Measurement*, 45(4), 343-362.
- de La Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69(3), 333-353.
- de La Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis analysis of fraction subtraction data. *Psychometrika*, 73(4), 595-624.
- DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. En P. D. Nichols, S. F. Chipman & R. L. Brennan (Eds.), *Cognitively diagnostic assessment*. (pp. 361-389). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.
- Dimitrov, D. M. (2007). Least squares distance method of cognitive validation and analysis for binary items using their item response theory parameters. *Applied Psychological Measurement*, 31(5), 367-387.
- Dimitrov, D. M., & Raykov, T. (2003). Validation of cognitive structures: A structural equation modeling approach. *Multivariate Behavioral Research*, 38(1), 1-23.
- Dimitrov, D. M., Romero, S. J., Ponsoda, V., & Ximénez, C. (2006, October). *Psychometric analysis of cognitive operations underlying student performance on basic arithmetic operations: An application of the Least Squares Distance Method*. Paper presented at the 2006 Annual Meeting of the Mid-Western Educational Research Association (MWERA), Columbus, Ohio.
- Dimitrov, D. M., Romero, S. J., Ponsoda, V., & Ximénez, C. (2007, April). *An IRT based analysis of arithmetic test items*. Paper presented at the 2007 AERA Annual Meeting, Chicago, Illinois.

- Dixon, J. A., Deets, J. K., & Bangert, A. (2001). The representations of the arithmetic operations include functional relationships. *Memory & Cognition*, 29(3), 462-477.
- Embretson, S. (1984). A general latent trait model for response processes. *Psychometrika*, 49(2), 175-186.
- Embretson, S. (1993). Psychometric models for learning and cognitive processes. En N. Frederiksen, R. J. Mislevy & I. I. Bejar (Eds.), *Test theory for a new generation of tests*. (pp. 125-150). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.
- Embretson, S. E. (1985). Multicomponent latent trait models for test design. En S. E. Embretson (Ed.), *Test design. developments in psychology and psychometrics* (pp. 195-196-218). New York: Academic Press, Inc.
- Embretson, S. E. (1996). Multidimensional latent trait models in measuring fundamental aspects of intelligence. En I. Dennis, & P. Tapsfield (Eds.), *Human abilities: Their nature and measurement*. (pp. 117-132). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.
- Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, 3(3), 380-396.
- Embretson, S. E. (1999a). Cognitive psychology applied to testing. In F. T. Durso (Ed.), *Handbook of applied cognition*. (pp. 629-660). New York, NY, US: John Wiley & Sons Ltd.
- Embretson, S. E. (1999b). Generating items during testing: Psychometric issues and models. *Psychometrika*, 64(4), 407-433.
- Embretson, S. E., & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement*, 11(2), 175-193.
- Embretson, S. E., & Yang, X. (2006). Multicomponent latent trait models for complex tasks. *Journal of Applied Measurement*, 7(3), 335-350.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861-862-864.
- Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37(6), 359-374.
- Fischer, G. H. (1995). The linear logistic test model. En G. H. Fisher, & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications*. New York: Springer-Verlag.

- Fischer, G. H., & Ponocny-Seliger, E. (1998). *Structural Rasch modeling. handbook of the usage of PLCM-WIN 1.0*. Groningen, the Netherlands: proGAMMA.
- Fischer, G. H., & Tanzer, N. K. (1994). Some LBTL and LLTM relationships. In G. H. Fisher, & D. Laming (Eds.), *Contributions to mathematical psychology, psychometrics and methodology* (pp. 278-279-303). New York: Springer-Verlag.
- Fischer, M. H. (2003). Spatial representations in number processing: evidence from a pointing task. *Visual Cognition*, 10(4), 493-508.
- Franco, M., & Vivo, J. M. (2007). *Análisis de curvas ROC: principios básicos y aplicaciones*. Madrid: La Muralla.
- Fraser, C. & McDonald, R.P. (1988). NOHARM: least squares item factor analysis. *Multivariate Behavioural Research*, 23, 267-269.
- Fuson, K. C. (1992). Research on whole number addition and subtraction. En D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning: A project of the national council of teachers of mathematics*. (pp. 243-275). New York, NY, England: Macmillan Publishing Co, Inc.
- Gallistel, C. R. & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition*, 44, 43-74.
- Gitomer, D. H., Steinberg, L. S., & Mislevy, R. J. (1995). Diagnostic assessment of troubleshooting skill in an intelligent tutoring system. En P. D. Nichols, S. F. Chipman & R. L. Brennan (Eds.), *Cognitively diagnostic assessment*. (pp. 73-101). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.
- Gorin, J. S. (2007). Test construction and diagnostic testing. En J. P. Leighton, & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications*. (pp. 173-201). New York, NY, US: Cambridge University Press.
- Greeno, J. G., Collins, A. M., & Resnick, L. B. (1996). Cognition and learning. En D. C. Berliner, & R. C. Calfee (Eds.), *Handbook of educational psychology*. (pp. 15-46). New York, NY: Prentice Hall International.
- Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, 26(4), 301-321.
- Hartz, S. M. (2002). *A bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. Unpublished doctoral dissertation, University of Illinois, Champaign.

- Hartz, S. M., Roussos, L., & Stout, W. (2002). *Skills diagnosis: Theory and practice (user manual for arpeggio software)*. Princeton, NJ: Educational Testing Service.
- Hegarty, M., Mayer, R. E., & Monk, C. A. (1995). Comprehension of arithmetic word problems: A comparison of successful and unsuccessful problem solvers. *Journal of Educational Psychology*, 87(1), 18-32.
- Hogaboam, T. W., & Pellegrino, J. W. (1978). Hunting for individual differences in cognitive processes: Verbal ability and semantic processing of pictures and words. *Memory & Cognition*, 6(2), 189-193.
- Hornke, L. F., & Habon, M. W. (1986). Rule-based item bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement. Special Issue: Test Item Banking*, 10(4), 369-380.
- Johnson, P. J., Goldsmith, T. E., & Teague, K. W. (1995). Similarity, structure, and knowledge: A representational approach to assessment. En P. D. Nichols, S. F. Chipman & R. L. Brennan (Eds.), *Cognitively diagnostic assessment*. (pp. 221-249). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.
- Junker, B.W., y Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3), 258-272.
- Kameenui, E. J., & Griffin, C. C. (1989). The national crisis in verbal problem solving in mathematics: A proposal for examining the role of basal mathematics programs. *The Elementary School Journal*, 89(5), 575-593.
- Kamii, C., & Joseph, L. L. (1989). *Young children continue to reinvent arithmetic--2nd grade: Implications of piaget's theory*. New York, NY, US: Teachers College Press.
- Karelitz, T. M. (2008). How binary skills obscure the transition from non-mastery to mastery. *Measurement: Interdisciplinary Research and Perspectives*, 6(4), 268-272.
- Lawson, C. L., & Hanson, R. J. (1974). *Solving least squares problems*. Englewood Cliffs, New Jersey: Prentice-Hall.
- Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.
- LeFevre, J., Sadesky, G. S., & Bisanz, J. (1996). Selection of procedures in mental addition: Reassessing the problem size effect in adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(1), 216-230.

- Leighton, J. P. & Gierl, M. J. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. New York, NY: Cambridge University Press.
- Lohman, D. F. (1989). Individual differences in errors and latencies on cognitive tasks. *Learning and Individual Differences. Special Issue: Methodology and Individual Differences*, 1(2), 179-202.
- Lohman, D. F. (2000). Complex information processing and intelligence. In R. J. Sternberg (Ed.), *Handbook of intelligence*. (pp. 285-340). New York, NY: Cambridge University Press.
- López, A., & Elousa, P. (2002). Formulación y validación de un modelo logístico lineal para la tarea de adición y sustracción de fracciones y números mixtos. *Psicothema*, 14(4), 802-809.
- Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Oxford, England: Addison-Wesley.
- Ma, L. Çetin, E. & Green, K. E. (2009, April). *Cognitive assessment in Mathematics with the Least Squares Distance Method*. Paper presented at the 2009 AERA Annual Meeting, San Diego, California.
- Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*, 2(2), 99-120.
- Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, 64(2), 187-212.
- Marshall, S. P. (1993). Assessing schema knowledge. En N. Frederiksen, R. J. Mislevy & I. I. Bejar (Eds.), *Test theory for a new generation of tests*. (pp. 155-180). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174.
- Maza, C. (1991). *Enseñanza de la suma y de la resta*. Madrid: Síntesis.
- Maza, C. (2001). Adición y Sustracción. En E. de Castro (Ed.), *Didáctica de la Matemática en la Educación Primaria* (pp. 177-178-202). Madrid: Síntesis.
- Medina-Díaz, M. (1993). Analysis of cognitive structure using the linear logistic test model and quadratic assignment. *Applied Psychological Measurement*, 17(2), 117-130.

- Messick, S. (1989). Validity. En R. L. Linn (Ed.), *Educational measurement (3rd ed.)*. (pp. 13-103). New York, NY, England: Macmillan Publishing Co, Inc.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749.
- Miaralet, G. (1984). *Las matemáticas: cómo se aprenden, cómo se enseñan?*. Madrid: Visor.
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2), 81-97.
- Mislevy, R. J. (1995). Probability-based inference in cognitive diagnosis. En P. D. Nichols, S. F. Chipman & R. L. Brennan (Eds.), *Cognitively diagnostic assessment*. (pp. 43-71). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.
- Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, 55(2), 195-215.
- Mislevy, R. J., Almond, R. G., Yan, D., & Steinberg, L. S. (2000). *Bayes nets in educational assessment: Where do the numbers come from? CSE technical report* No. CSE-TR-518.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). *On the structure of educational assessments, CSE technical report* No. CSE-TR-597.
- Mislevy, R. J., Steinberg, L. S., Almond, R. G., Breyer, F. J., & Johnson, L. (2001). *Making sense of data from complex assessments. CSE technical report* No. CSE-TR-538.
- Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference*. Thousand Oaks, CA, US: Sage Publications, Inc.
- Muldoon, K. P., Lewis, C., & Francis, B. (2007). Using cardinality to compare quantities: The role of social-cognitive conflict in early numeracy. *Developmental Science*, 10(5), 694-711.
- Naveh-Benjamin, M., Lin, Y., & McKeachie, W. J. (1995). Inferring students' cognitive structures and their development using the "fill-in-the-structure" (FITS) technique. En P. D. Nichols, S. F. Chipman & R. L. Brennan (Eds.), *Cognitively diagnostic assessment*. (pp. 279-304). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

- Nichols, P. D. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research*, 64(4), 575-603.
- Nichols, P. D., Chipman, S. F., & Brennan, R. L. (1995). *Cognitively diagnostic assessment*. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.
- Penner-Wilger, M., Leth-Steensen, C., & LeFevre, J. (2002). Decomposing the problem-size effect: A comparison of response time distributions across cultures. *Memory & Cognition*, 30(7), 1160-1167.
- Peterson, M. J., & Aller, S. (1971). Arithmetic problem solving. *Journal of Experimental Psychology*, 91(1), 93-97.
- Real, E., Olea, J., Ponsoda, V., Revuelta, J., & Abad, F. J. (1999). Análisis de la dificultad de un test de matemáticas mediante un modelo componencial. *Psicológica*, 20(2), 121-134.
- Rizopoulos, D. (2006). Ltm: An R package for latent variable modeling and item response analysis. *Journal of Statistical Software*, 17(5) Recuperado de <http://www.jstatsoft.org/v17/i05>
- Romero, S. J., Ordoñez, X., López, E., & Navarro, E. (2009). Análisis de la estructura cognitiva de la competencia científica en PISA 2006 mediante el LSDM: el caso español. *Psicothema*, 21, 509-514.
- Romero, S. J., Ponsoda, V., & Ximénez, C. (2006). Validación de la estructura cognitiva del test de signos mediante modelos de ecuaciones estructurales. *Psicothema*, 18(4), 835-840.
- Romero, S. J., Ponsoda, V., & Ximénez, C. (2008). Análisis de un test de aritmética mediante el modelo logístico lineal de rasgo latente. *Revista Latinoamericana De Psicología*, 40(1), 85-95.
- Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14(3), 271-282.
- Roussos, L. A., DiBello, L. V., Stout, W., Hartz, S. M., Henson, R. A., & Templin, J. L. (2007). The fusion model skills diagnosis system. En J. P. Leighton, & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications*. (pp. 275-318). New York, NY, US: Cambridge University Press.
- Roussos, L., & Karelitz, T. (2004). *The effects of attribute heterogeneity on the performance of arpeggio and GAMEstats* (Unpublished ETS Project Report). Princeton, NJ: Recuperado de

http://www.measuredprogress.org/resources/psychometrics/framework/materials/03_06/EffectsAttribute.pdf

- Rupp, A. A. (2007). The answer is in the question: A guide for describing and investigating the conceptual foundations and statistical properties of cognitive psychometric models. *International Journal of Testing*, 7(2), 95-125.
- Rupp, A. A., & Mislevy, R. J. (2007). Cognitive foundations of structured item response models. En J. P. Leighton, & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications*. (pp. 205-241). New York, NY, US: Cambridge University Press.
- Rupp, A. A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement*, 68(1), 78-96.
- Schwarz, B. B., Kohn, A. S., & Resnick, L. B. (1993-1994). Positives about negatives: A case study of an intermediate model for signed numbers. *Journal of the Learning Sciences*, 3(1), 37-92.
- Seyler, D. J., Kirk, E. P., & Ashcraft, M. H. (2003). Elementary subtraction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(6), 1339-1352.
- Sinharay, S., Almond, R., & Yan, D. (2004). *Assessing fit of models with discrete proficiency variable in educational assessment*. ETS research report. N° RR-04.
- Smits, D. J. M., & De Boeck, P. (2003). A componential IRT model for guilt. *Multivariate Behavioral Research*, 38(2), 161-188.
- Smits, D. J. M., De Boeck, P., & Verhelst, N. D. (2003). Estimation of the MIRID: A program and a SAS-based approach. *Behavior Research Methods, Instruments & Computers*, 35(4), 537-549.
- Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. En R. L. Linn (Ed.), *Educational measurement (3rd ed.)*. (pp. 263-331). New York, NY, England: Macmillan Publishing Co, Inc.
- Spada, H. (1977). Logistic models of learning and thought. En H. Spada, & W. F. Kempf (Eds.), *Structural models of thinking and learning* (pp. 227-262)
- Sternberg, R. J. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities*. Oxford, England: Lawrence Erlbaum.

- Sternberg, R. J., & Weil, E. M. (1980). An aptitude X strategy interaction in linear syllogistic reasoning. *Journal of Educational Psychology*, 72(2), 226-239.
- Suppes, P. (1969). Stimulus-response theory of finite automata. *Journal of Mathematical Psychology*, 6(3), 327-355.
- Tanaka, J. S., & Huba, G. J. (1985). A fit index for covariance structure models under arbitrary GLS estimation. *British Journal of Mathematical and Statistical Psychology*, 38(2), 197-201.
- Tanzer, N. K., Gittler, G., & Ellis, B. B. (1995). Cross-cultural validation of item complexity in a LLTM-calibrated spatial ability test. *European Journal of Psychological Assessment. Special Issue: Testing Across Cultures: Theoretical Issues and Empirical Results*, 11(3), 170-183.
- Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20(4), 345-354.
- Tatsuoka, K. K. (1984a). Caution indices based on item response theory. *Psychometrika*, 49(1), 95-110.
- Tatsuoka, K. K. (1984b). Changes in error types over learning stages. *Journal of Educational Psychology*, 76(1), 120-129.
- Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics*, 10(1), 55-73.
- Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. En P. D. Nichols, S. F. Chipman & R. L. Brennan (Eds.), *Cognitively diagnostic assessment*. (pp. 327-359). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.
- Tatsuoka, K. K., & Boodoo, G. M. (2000). Subgroup differences on the GRE quantitative test based on the underlying cognitive processes and knowledge. En: A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in Mathematics and Science*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Tatsuoka, K. K., & Tatsuoka, M. M. (1987). Bug distribution and statistical pattern classification. *Psychometrika*, 52(2), 193-206.
- Tatsuoka, K. K., Corter, J. E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal*, 41(4), 901-926.

- van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.
- van Lehn, K. (1989). Problem solving and cognitive skill acquisition. En M. I. Posner (Ed.), *Foundations of cognitive science*. (pp. 527-579). Cambridge, MA, US: The MIT Press.
- von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, 61(2), 287-307.
- von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial-credit model. *Applied Psychological Measurement*, 28(6), 389-406.
- White, B. Y., & Frederiksen, J. R. (1987). Qualitative models and intelligent learning environments. En R. W. Lawler, & M. Yazdani (Eds.), *Artificial intelligence and education, vol. 1: Learning environments and tutoring systems*. (pp. 281-305). Westport, CT, US: Ablex Publishing.
- Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika*, 45(4), 479-494.
- Whitely, S. E., & Schneider, L. M. (1981). Information structure for geometric analogies: A test theory approach. *Applied Psychological Measurement*, 5(3), 383-397.
- Xin, T., Xu, Z., & Tatsuoka, K. (2004). Linkage between teacher quality, student achievement, and cognitive skills: A rule-space model. *Studies in Educational Evaluation*, 30(3), 205-223.
- Yamamoto, K., & Gitomer, D. H. (1993). Application of a HYBRID model to a test of cognitive skill representation. En N. Frederiksen, R. J. Mislevy & I. I. Bejar (Eds.), *Test theory for a new generation of tests*. (pp. 275-295). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.
- Yang, X., & Embretson, S. E. (2007). Construct validity and cognitive diagnostic assessment. En J. P. Leighton, & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications*. (pp. 119-145). New York, NY, US: Cambridge University Press.
- Yip, M. C. W. (2002). Presentation effects on arithmetic problem solving. *Psychologia: An International Journal of Psychology in the Orient*, 45(2), 90-97.